

# HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

Fourth Edition



Edited by

**Gerald Goldstein<sup>(†)</sup>**  
**Daniel N. Allen**  
**John DeLuca**



# **Handbook of Psychological Assessment**

# **Handbook of Psychological Assessment**

**Fourth Edition**

*Edited by*

**Gerald Goldstein<sup>†</sup>**

**VA Pittsburgh Healthcare System,  
Pittsburgh, PA, United States**

**University of Pittsburgh, Pittsburgh,  
PA, United States**

**Daniel N. Allen**

**Department of Psychology, University of  
Nevada, Las Vegas, NV, United States**

**John DeLuca**

**Kessler Foundation, West Orange,  
NJ, United States**

<sup>†</sup>Deceased



**ACADEMIC PRESS**

An imprint of Elsevier

Academic Press is an imprint of Elsevier  
125 London Wall, London EC2Y 5AS, United Kingdom  
525 B Street, Suite 1650, San Diego, CA 92101, United States  
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States  
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2019 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: [www.elsevier.com/permissions](http://www.elsevier.com/permissions).

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

#### Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

#### British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

#### Library of Congress Cataloguing-in-Publication Data

A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-802203-0

For Information on all Academic Press publications  
visit our website at <https://www.elsevier.com/books-and-journals>



Working together  
to grow libraries in  
developing countries

[www.elsevier.com](http://www.elsevier.com) • [www.bookaid.org](http://www.bookaid.org)

*Publisher:* Nikki Levy

*Acquisition Editor:* Nikki Levy

*Editorial Project Manager:* Barbara Makinster

*Production Project Manager:* Vijayaraj Purushothaman

*Cover Designer:* Victoria Pearson

Typeset by MPS Limited, Chennai, India

# **Dedication**

Gerald Goldstein, September 16, 1931–April 8, 2017

This book is dedicated to Gerald “Jerry” Goldstein  
Scholar, Mentor, Friend

—DNA and JDL

# List of contributors

**Anna V. Agranovich** Department of Physical Medicine and Rehabilitation, Johns Hopkins University School of Medicine, Baltimore, MD, United States

**Daniel N. Allen** Department of Psychology, University of Nevada, Las Vegas, NV, United States

**Teresa A. Ashman** Private Practice, New York, NY, United States; NYU Langone Medical Center, New York, NY, United States

**Victoria Bacon** Department of Psychology, University of Nevada, Las Vegas, NV, United States

**Megan L. Becker** Department of Psychology, University of Nevada, Las Vegas, NV, United States

**Franklin C. Brown** Yale University, Department of Neurology, New Haven, CT, United States

**James N. Butcher** Emeritus Professor, University of Minnesota, Minneapolis, MN, United States

**Yen-Ling Chen** University of Nevada, Las Vegas, LV, United States

**Karen L. Dahlman** Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, United States

**John DeLuca** Kessler Foundation Research Center, West Orange, NJ, United States

**Ruben J. Echemendia** UOC Concussion Care Clinic, State College, Pennsylvania, PA, United States; University of Missouri-Kansas City, Pennsylvania, PA, United States

**Philip Erdberg** University of California - San Francisco, CA, United States

**Andrew Freeman** Department of Psychology, University of Nevada, Las Vegas, NV, United States

**Andrew J. Freeman** University of Nevada, Las Vegas, NV, United States

**Gerald Goldstein** VA Pittsburgh Healthcare System, Pittsburgh, PA, United States

**Gabriela González** Florida Institute of Technology, Melbourne, FL, United States

**Ross W. Greene** Department of Psychology, Virginia Tech, Blacksburg, VA, United States

**Jo-Ida C. Hansen** Center for Interest Measurement Research, Department of Psychology, University of Minnesota, Minneapolis, MN, United States

**Stephen N. Haynes** Department of Psychology, University of Hawai'i at Mānoa, Honolulu, HI, United States

**James A. Holdnack** Research and Statistics Consultant, Bear, DE, United States

**Arthur MacNeill Horton, Jr.** Psych Associates of Maryland, Towson, Columbia, MD, United States

**Joseph Keawe‘aimoku Kaholokula** Department of Native Hawaiian Health, John A. Burns School of Medicine, University of Hawai'i at Mānoa, Honolulu, HI, United States

**Robert L. Kane** Cognitive Consults and Technology LLC, Washington Neuropsychology Research Group, Neuropsychology Associates of Fairfax, Fairfax, VA, United States

**Lynda J. Katz** Private Practice, Durham, NC, United States

**Christopher A. Kearney** Department of Psychology, University of Nevada, Las Vegas, NV, United States

**Mary Lynne Kennedy** Neuropsychology Partners Incorporated, East Providence, RI, United States

**Jeannie Lengenfelder** Department of Physical Medicine & Rehabilitation, Rutgers, the State University of New Jersey-New Jersey Medical School, Newark, NJ, United States

**Ron Livingston** The University of Texas at Tyler, Tyler, TX, United States

**Zarui A. Melikyan** University of California Irvine, Institute for Memory Impairments and Neurological Disorders, Irvine, CA, United States

**Victoria C. Merritt** VA San Diego Healthcare System, San Diego, CA, United States

**Richard C. Mohs** Global Alzheimer's Platform Foundation, Washington, DC, United States

**William H. O'Brien** Department of Psychology, Bowling Green State University, Bowling Green, OH, United States

**Thomas H. Ollendick** Department of Psychology, Virginia Tech, Blacksburg, VA, United States

**Thomas D. Parsons** Director, NetDragon Digital Research Centre, University of North Texas, Denton, TX, United States; Director, Computational Neuropsychology and Simulation (CNS) Lab, Department of Psychology, University of North Texas, Denton, TX, United States; Professor, College of Information, University of North Texas, Denton, TX, United States

**Jacob A. Paulsen** San Diego City College, San Diego, CA, United States

**Antonio E. Puente** University of North Carolina Wilmington, Department of Psychology, Wilmington, NC, United States

**Cecil Reynolds** Texas A&M University, Austin, TX, United States

**Henry V. Soper** Fielding Graduate University, Santa Barbara, CA, United States

**Frank M. Webbe** Florida Institute of Technology, Melbourne, FL, United States

**Michael D. Weiler** Cranston Public School System, Cranston, RI, United States

**Carolyn L. Williams** Emeritus Professor, University of Minnesota, Minneapolis, MN, United States

**W. Grant Willis** University of Rhode Island, Kingston, RI, United States

# Preface

Since the publication of the third edition of this handbook in 2000 there have been many new developments in the field of psychological testing and assessment. The use of technology has increased, new versions of standard tests like the Wechsler Intelligence scales were published, and neuropsychological assessment research continues to appear at a rapid pace. These and other areas reflecting new developments are addressed in entirely new chapters in the Handbook. Some of the authors who wrote chapters previously have revised their work in a manner that reflects the significant developments in their specialties since the third edition was published. Consistent with prior editions, we have continued the practice of inviting new authors to write some of the chapters in order to provide different perspectives and theoretical frameworks from authorities in their areas. We attempted as much as possible to preserve the book as a basic reference work, while also providing current information. The editors would like to thank the new authors for offering their new perspectives and philosophies of assessment, and the authors who wrote chapters previously for their conscientious and detailed updates.

# **Part I**

## **Introduction**

# Historical perspectives

1

Gerald Goldstein<sup>1</sup>, Daniel N. Allen<sup>2</sup> and John DeLuca<sup>3</sup>

<sup>1</sup>VA Pittsburgh Healthcare System, Pittsburgh, PA, United States, <sup>2</sup>Department of Psychology, University of Nevada, Las Vegas, NV, United States, <sup>3</sup>Kessler Foundation Research Center, West Orange, NJ, United States

## Introduction

The first edition of the *Handbook of Psychological Assessment* was published with Michel Hersen in 1984, followed by a second edition in 1990 and a third in 1999. In this chapter of the present edition we reproduce the historical material presented earlier with some modifications and provide an update of historical developments during the almost 20 years since the appearance of the last edition. We began the first edition indicating that “A test is a systematic procedure for comparing the behavior of two or more persons.” This definition of a test, offered by Lee Cronbach many years ago (1949/1960) probably still epitomizes the major content of psychological assessment. The invention of psychological tests, then known as mental tests, is generally attributed to Galton (Boring, 1950), and occurred during the middle to late 19th century. Galton’s work was largely concerned with differences between individuals, and his approach was essentially in opposition to the approaches of other psychologists of his time. Most psychologists then were primarily concerned with the exhaustive study of mental phenomena in a few participants, while Galton was more interested in somewhat less specific analyses of large numbers of people. Perhaps the first psychological test was the “Galton whistle,” which evaluated high tone hearing. Galton also appeared to have believed in the statistical concept that held that errors of measurement in individuals could be canceled out through the mass effect of large samples. Obviously, psychologists have come a long way from the simple tests of Galton, Binet, and Munsterberg, and the technology of testing is now in the computer age, with almost science fiction—like extensions, such as testing by satellite and virtual reality applications. Psychometrics is now an advanced branch of mathematical and statistical science, and the administration, scoring, and even interpretation of tests has become increasingly objectified and automated. There are numerous books and scientific journals now devoted entirely to assessment. While some greet the news with dread and others with enthusiasm, we may be rapidly approaching the day when most of all testing will be administered, scored, and interpreted by computer. Thus, the 19th-century image of the school teacher administering paper-and-pencil tests to the students in the

classroom and grading them at home has given way to the extensive use of automated procedures administered to large portions of the population by representatives of large corporations. Testing appears to have become a part of western culture, and there are indeed very few people who enter educational, work, or clinical settings who do not take many tests during their lifetimes. The presence of testing laboratories equipped with computers in clinical and educational settings is now not uncommon.

In recent years, there has been a distinction made between testing and assessment, assessment being the broader concept. Psychologists do not just give tests now; they perform assessments. The title of this volume, the *Handbook of Psychological Assessment*, was chosen advisedly and is meant to convey the view that it is not simply a handbook of psychological testing, although testing is covered in great detail. The term assessment implies that there are many ways of evaluating individual differences. Testing is one way, but there are also others, including interviewing, observations of behavior in natural or structured settings, and the recording of various physiological functions. Certain forms of interviewing and systematic observation of behavior are now known as behavioral assessments, as opposed to the psychometric assessment accomplished through the use of formal tests. Historically, interest in these two forms of assessment has waxed and waned, and in what follows we will briefly try to trace these trends in various areas.

## **Intelligence and achievement testing**

The testing of intelligence in school children was probably the first major occupation of clinical psychology. However, advocacy efforts to improve the quality of education, such as the “No Child Left Behind” federal program, have become associated with substantially increased use of testing of academic abilities. Tests of reading, writing, and mathematical abilities have become increasingly used over the past several decades. The Binet scales and their descendants continue to be used, along with the IQ concept associated with them. Later, primarily through the work of David Wechsler and associates (Wechsler, 1944), intelligence testing was extended to adults and the IQ concept was changed from the mental age system (Mental Age/Chronological Age  $\times 100$ ) to the notion of a deviation IQ based on established population-based norms. While Wechsler was primarily concerned with the individual assessment of intelligence, many group-administered paper-and-pencil tests also emerged during the early years of the 20th century. These tests were generally designed to allow for the assessment of large groups of individuals in situations where individually administered tests were impractical. The old Army Alpha and Beta tests, developed for the intellectual screening of large groups of inductees into the armed forces during the First World War, were among the first examples of these instruments. In recent times efficiencies afforded by group testing have been assisted by self-administration of tests using computers.

Use of these tests progressed in parallel with developments in more theoretical research regarding the nature of intelligence. The English investigators Burt, Pearson, and Spearman and the Americans Thurstone and Guilford are widely known for their work in this area, particularly with factor analysis. The debate over whether intelligence is a general ability (g) or a series of specific abilities represents one of the classic controversies in psychology. Factor analysis has provided support for various models describing the structure of intelligence, and has expanded early conceptualizations of intelligence as composed primarily of verbal and nonverbal abilities or fluid and crystallized abilities to the more complex models now used to describe the structure of the Wechsler scales and hierarchies of general, broad, and narrow abilities that typify the Cattell–Horn–Carroll theory ([McGrew, 2005](#)).

Another highly significant aspect of intelligence testing has to do with its clinical utilization. The IQ now essentially defines the borders of intellectual ability and disability, formerly called mental retardation, and intelligence tests are widely used to identify disabled children in educational settings [American Psychiatric Association (APA), Diagnostic and Statistical Manual of Mental Disorders, 5th ed., ([DSM-5, 2013](#))]. However, intelligence testing has gone far beyond the attempt to identify intellectually disabled individuals and has become widely applied in the fields of psychopathology and neuropsychology. With regard to psychopathology, under the original impetus of David Rapaport and collaborators ([Rapaport, Gill, & Schafer, 1945](#)), the Wechsler scales became clinical instruments used in conjunction with other tests to evaluate patients with such conditions as schizophrenia and various stress-related disorders. In the field of neuropsychology, use of intelligence testing is possibly best described by [McFie's \(1975\)](#) remark, “It is perhaps a matter of luck that many of the Wechsler subtests are neurologically relevant” (p. 14). In these applications, the intelligence test was basically used as an instrument by the clinician to examine various cognitive processes in order to make inferences about the patient’s clinical status. In summary, the intelligence test has become a widely used assessment instrument in educational, industrial, military, and clinical settings. While in some applications the emphasis remains on the simple obtaining of a numerical IQ value, it is probably fair to say that many, if not most, psychologists now use the intelligence test as a means of examining the individual’s cognitive processes; of seeing how he or she goes about solving problems; of identifying those factors that may be interfering with adaptive thinking and behavior; of looking at various language and nonverbal abilities in brain-damaged patients; and of identifying patterns of abnormal thought processes seen in schizophrenia, autism, and other patient groups. As the theoretical models proposed to understand IQ have become increasingly complex there has been an accompanying increase in the development of various index scores to reflect performance on current versions of intelligence tests. Performance profiles and qualitative characteristics of individual responses to items appear to have become the major foci of interest, rather than the single IQ score. The recent appearance of the new child and adult versions of the Wechsler intelligence scales reflect the major impacts cognitive psychology and neuropsychology have had on the way in which intelligence is currently conceptualized and intelligence test results are currently interpreted.

## Personality assessment

Personality assessment has come to rival intelligence testing as a task performed by psychologists. However, while most psychologists would agree that an intelligence test is generally the best way to measure intelligence, no such consensus exists for personality evaluation. From a long-term perspective, it would appear that two major philosophies and perhaps three assessment methods have emerged. The two philosophies can be traced back to Allport's (1937) distinction between nomothetic versus idiographic methodologies and Meehl's (1954) distinction between clinical and statistical or actuarial prediction. In essence, some psychologists feel that personality assessments are best accomplished when they are highly individualized, while others have a preference for quantitative procedures based on group norms. The phrase "seer versus sign" coined in a paper by Lindzey (1965) has been used to epitomize the dispute regarding whether the judgment of clinicians (seer) or actuarial statistical approaches (sign) provide superior predictive efficiency. The three methods referred to are the interview and the projective and objective tests.

### ***The interview***

Obviously, the initial way that psychologists and their predecessors found out about people was to talk to them, giving the interview historical precedence. Following a period wherein the use of the interview was eschewed by many psychologists, it now has made a return. It would appear that the field is in a historical spiral, with various methods leaving and returning at different levels. The interview began as a relatively unstructured conversation with the patient and perhaps an informant, with varying goals including obtaining a history, assessing personality structure and dynamics, establishing a diagnosis, and many other matters. Numerous publications have been written about interviewing (e.g., Menninger, 1952), but in general they provided outlines and general guidelines as to what should be accomplished by the interview. However, model interviews were not provided. With or without this guidance, the interview was viewed by many as a subjective, unreliable procedure that could not be sufficiently validated. For example, the unreliability of psychiatric diagnosis based on studies of multiple interviewers had been well established (Zubin, 1967). In reaction to unreliability of clinical interviews at the time, several structured psychiatric interviews appeared in which the specific content, if not specific items, is presented, and for which very adequate reliability has been established. There are now several such interviews available, including the Schedule for Affective Disorders and Schizophrenia (SADS) (Spitzer & Endicott, 1977), the Renard Diagnostic Interview (Helzer, Robins, Croughan, & Welner, 1981), and the Structured Clinical Interview for DSM-III, DSM-III-R, or DSM-IV (SCID or SCID-R) (Spitzer & Williams, 1983) (now updated for DSM-5). These interviews have been established in conjunction with objective diagnostic criteria included in the DSMs, the Research Diagnostic Criteria (Spitzer, Endicott, & Robins, 1977), and the Feighner Criteria (Feighner et al., 1972). Some have been adapted for

application in both research and clinical settings. These procedures have apparently ushered in a “comeback” of the interview, and many psychiatrists and psychologists now prefer to use these procedures when making psychiatric diagnoses rather than either the objective or projective-type psychological test. Those advocating use of structured interviews point to the fact that in psychiatry, at least, tests must ultimately be validated against diagnostic judgments made by psychiatrists. These judgments are generally based on interviews and observation, since there really are no specific biological or other objective markers of most forms of psychopathology. If that is indeed the case, there seems little point in administering elaborate and often lengthy tests when one can just as well use the criterion measure itself, that is, the interview, rather than the test. Put another way, there is no way that a test can be more valid than an interview if an interview is the ultimate validating criterion.

Structured interviews have made a major impact on the scientific literature in psychopathology, and it is rare to find a recently written research report in which the diagnoses were not established by one of them. It would appear that we have come full cycle regarding this matter, and until objective and specific markers of various forms of psychopathology are discovered, we will continue to rely primarily on the structured interviews for diagnostic assessments. Interviews such as the SCID or the Diagnostic Interview Schedule (DIS) type are relatively lengthy and comprehensive, but there are also several briefer, more specific interview or interview-like procedures that allow for diagnosis of specific conditions, as well as procedures that allow clinicians to make severity ratings of specific symptoms based on an interview with the client. Within psychiatry, perhaps the most well-known procedure is the Brief Psychiatric Rating Scale (BPRS) ([Overall & Gorham, 1962](#)). In the area of affective disorders, the Hamilton Depression Scale ([Hamilton, 1960](#)) and Young Mania Rating Scale ([Young, Biggs, Ziegler, & Meyer, 1978](#)) have played similar roles historically. A wide range of specific interview procedures are also available for examination of psychotic symptoms including the positive and negative symptom scales developed by [Andreasen \(1984\)](#) as well as newer second generation measures developed to assess negative symptoms based on more recent theoretical models ([Kirkpatrick et al., 2011; Kring, Gur, Blanchard, Horan, & Reise, 2013](#)). Many other interview based procedures are available as well. Unlike the SCID and similar measures, interview procedures like the BPRS are often not tied directly to DSM or other diagnostic criteria. However, they do allow for examination of presence and severity of symptoms consistent with specific diagnoses, and when repeated, reflect a standardized way to examine change in patient symptoms, usually as a function of taking some form of psychotropic medication or in response to behavioral interventions.

There are also several widely used interviews for patients with dementia, which generally combine a brief mental status examination and some form of functional assessment, with particular reference to activities of daily living. Historically, the most popular of these scales are the Mini-Mental Status Examination of [Folstein, Folstein, and McHugh \(1975\)](#), and the Dementia Scale of [Blessed, Tomlinson, and Roth \(1968\)](#). Extensive validation studies have been conducted with these

instruments, perhaps the most well-known study having to do with the correlation between scores on the Blessed, Tomlinson, and Roth scale used in patients while they are living and the senile plaque count determined on autopsy in patients with dementia. The obtained correlation of 0.7 quite impressively suggested that the scale was a valid one for detection of dementia. Since the publication of the last edition, the Montreal Cognitive Assessment has become one of the most commonly used interviews for patients with dementia (Ozer, Young, Champ, & Burke, 2016).

In addition to these interviews and rating scales, numerous methods have been developed by nurses and psychiatric aids for assessment of psychopathology based on direct observation of ward behavior (Raskin, 1982). Historically, the most widely used of these rating scales are the Nurses' Observation Scale for Inpatient Evaluation (NOSIE-30) (Honigfeld & Klett, 1965) and the Ward Behavior Inventory (Burdock, Hardesty, Hakerem, Zubin, & Beck, 1968). These scales assess such behaviors as cooperativeness, appearance, communication, aggressive episodes, and related behaviors, and are based on direct observation rather than reference to medical records or the reports of others. Scales of this type supplement the interview with information concerning social competence and the capacity to carry out functional activities of daily living. In recent years a new procedure called motivational interviewing has appeared, originally for use with problem drinkers; but it is really more of a counseling method than an assessment procedure used to help clients explore and resolve ambivalence about addressing their psychological concerns and motivating them for treatment (Rollnick & Miller, 1995).

Again taking a long-term historical view, it is our impression that after many years of neglect by the field, the interview has made a successful return to the arena of psychological assessment, but the interviews now used are quite different from the loosely organized, "freewheeling," conversation-like interviews of the past (Hersen & Van Hassett, 1998). First, their organization tends to be structured, and the interviewer is required to obtain certain items of information. Formulation of specifically worded questions is sometimes viewed as counterproductive but even in cases where this is the interview format, the interviewer, who should be an experienced clinician trained in the reliable use of the procedure, should be able to formulate questions that will elicit the required information. Second, the interview procedure must meet psychometric standards of validity and reliability. Finally, while structured interviews tend to be a theoretical in orientation, many are based on contemporary scientific knowledge of psychopathology. Thus, for example, the information needed to establish a differential diagnosis within the general classification of mood disorders is derived from current scientific literature on depression and related mood disorders as reflected in the DSM, or negative symptoms of psychosis are evaluated based on current theoretical understandings of symptom organization and structure.

### ***Projective personality tests***

The rise of the interview appears to have occurred in parallel with the decline of projective techniques. Those of us in a chronological category that may be roughly

described as elderly may recall that our graduate training in clinical psychology probably included extensive coursework and practicum experience involving the various projective techniques. Most clinical psychologists would probably agree that even though projective techniques are still used to some extent, the atmosphere of excitement concerning these procedures that existed during the 1940s and 1950s no longer seems to exist. Even though the Rorschach technique and Thematic Apperception Test (TAT) were the major procedures used during that era, a variety of other tests emerged quite rapidly: the projective use of human-figure drawings (Machover, 1949), the Szondi Test (Szondi, 1952), the Make-A-Picture-Story (MAPS) Test (Shneidman, 1952), the Four-Picture Test (VanLennep, 1951), the Sentence Completion Tests (e.g., Rohde, 1957), and the Holtzman Inkblot Test (Holtzman, 1958). The exciting work of Murray and his collaborators reported on in Explorations in Personality (Murray, 1938) had a major impact on the field and stimulated extensive utilization of the TAT. It would probably be fair to say that the sole survivor of this active movement is the Rorschach test. Some clinicians continue to use the Rorschach test, and the work of Exner and his collaborators has lent it increasing scientific respectability (see Dr. Philip Erdberg's Chapter 14: The Rorschach in this volume for a modern conceptualization of the Rorschach). Validity evidence now exists for a number of the Rorschach indices, particularly when used in the examination of individuals with psychotic disorders (Mihura, Meyer, Dumitrescu, & Bombel, 2013).

There are undoubtedly many reasons for the decline in utilization of projective techniques, but in our view they can be summarized by the three following points. First, increasing scientific sophistication created an atmosphere of skepticism concerning these instruments. Their validity and reliability were called into question by numerous studies (e.g., Swensen, 1957, 1968; Zubin, 1967), and a substantial segment of the professional community felt that the claims made for these procedures could not be substantiated. Second, developments in alternative procedures, notably the Minnesota Multiphasic Personality Inventory (MMPI) and other objective tests, convinced many clinicians that the information previously gained from projective tests could be gained more efficiently and less expensively with objective methods. In particular, the voluminous MMPI research literature demonstrated its usefulness in an extremely wide variety of clinical and research settings. When the MMPI and related objective techniques were pitted against projective techniques during the days of the "seer versus sign" controversy, it was generally demonstrated that sign was as good as or better than seer in most of the studies accomplished (Meehl, 1954). A current review of this research is contained in the chapter by Carolyn Williams, James Butcher, and Jacob Paulsen in this volume. Third, in general, the projective techniques are not atheoretical and, in fact, are generally viewed as being associated with one or another branch of psychoanalytic theory. While psychoanalysis remains a movement within psychology, there are numerous alternative theoretical systems at large, notably cognitive, behavioral, and biologically oriented systems. These alternative systems have largely supplanted psychoanalytic approaches and are consistent with the movement toward evidence-based interventions that have been proven effective as preferable to those with less validity.

evidence. As implied in the section of this chapter covering behavioral assessment, behaviorally oriented psychologists pose theoretical objections to projective techniques and make little use of them in their practices. Similarly, projective techniques tend not to currently receive high levels of acceptance in psychiatry departments which have become increasingly biologically oriented. In effect, then, utilization of projective techniques declined for scientific, practical, and philosophical reasons. However, the Rorschach test in particular continues to be used, often by psychodynamically oriented clinicians.

### ***Objective personality tests***

The early history of objective personality tests has been traced by [Cronbach \(1949, 1960\)](#). The beginnings apparently go back to Sir Francis Galton, who devised personality questionnaires during the latter part of the 19th century. We will not repeat that history here, but rather will focus on those procedures that survived into the contemporary era. In our view, there have been three such major survivors: a series of tests developed by Guilford and collaborators ([Guilford & Zimmerman, 1949](#)), a similar series developed by Cattell and collaborators ([Cattell, Eber, & Tatsuoka, 1970](#)), and the MMPI. In general, but certainly not in all cases, the Guilford and Cattell procedures are used for individuals functioning within the normal range, while the MMPI is more widely used in clinical populations. Thus, for example, Cattell's 16PF test may be used to screen job applicants, while the MMPI may be more typically used in psychiatric healthcare facilities. Furthermore, the Guilford and Cattell tests are based on factor analysis and are trait-oriented, while the MMPI in its original form did not make use of factor analytically derived scales and is more oriented toward psychiatric classification. Thus, the Guilford and Cattell scales contain measures of such traits as dominance or sociability, while most of the MMPI scales are named after psychiatric classifications such as paranoia or depression. Currently, most psychologists use one or more of these objective tests rather than interviews or projective tests in screening situations. For example, many thousands of patients admitted to psychiatric facilities operated by the Department of Veterans Affairs (VA) take the MMPI shortly after admission, while applicants for prison guard jobs in the state of Pennsylvania took the Cattell 16PF.

However, the MMPI in particular is commonly used as more than a screening instrument. It is frequently used as a part of an extensive diagnostic evaluation, as a method of evaluating treatment, and in numerous research applications. There is little question that it is the most widely used and extensively studied procedure in the objective personality-test area. Even though the 566 true-or-false items have remained essentially the same since the initial development of the instrument, the test has been revised and applications in clinical interpretation have evolved dramatically over the years. We have gone from perhaps an overly naive dependence on single-scale evaluations and overly literal interpretation of the names of the clinical scales (many of which are archaic psychiatric terms) to a sophisticated configurational interpretation of profiles, much of which is based on empirical research reviewed in the chapter by Dr. Williams and colleagues, and earlier by [Gilberstadt](#)

and Duker (1965) and Marks, Seeman, and Hailer (1974). Correspondingly, the methods of administering, scoring, and interpreting the MMPI have kept pace with technological and scientific advances in the behavioral sciences. From beginning with sorting cards into piles, hand scoring, and subjective interpretation, the MMPI has gone to computerized administration and scoring, interpretation based to a great extent on empirical research findings, and computerized interpretation. As is well known, there are several companies that will provide computerized scoring and interpretations of the MMPI. The MMPI has been completely revised and restandardized, and is now known as the MMPI-2. Since the appearance of the third edition of this handbook, use of the MMPI-2 has been widely adopted. Another procedure aside from the MMPI for objective personality assessment is presented in the work of Millon. Millon has produced a series of tests called the Millon Clinical Multiaxial Inventory (Versions I and II), the Millon Adolescent Personality Inventory, and the Millon Behavioral Health Inventory (Millon, 1982, 1985).

Even though we should anticipate continued spiraling of trends in personality assessment, it would appear that we have passed an era of projective techniques and are now living in a time of objective assessment, with an increasing interest in the structured interview. There also appears to be increasing concern with the scientific status of our assessment procedures. There has been particular concern about reliability of diagnosis, especially since distressing findings appeared in the literature suggesting that psychiatric diagnoses were being made quite unreliably (Zubin, 1967). The issue of validity in personality and psychopathology assessment remains a difficult one for a number of reasons. First, if by personality assessment we mean prediction or classification of some psychiatric diagnostic category, we have the problem of there being essentially no known objective markers for the major forms of psychopathology. Therefore, we were left essentially with psychiatrists' judgments. The more recent DSM systems have greatly improved this situation by providing objective criteria for the various mental disorders, but the capacity of such instruments as the MMPI or Rorschach test to predict DSM diagnoses remains an ongoing research question. Some scholars, however, have questioned the usefulness of taking that research course rather than developing increasingly reliable and valid structured interviews (Zubin, 1984). Similarly, there have been many reports of the failure of objective tests to predict such matters as success in an occupation or academic program, trustworthiness with regard to handling a weapon, and other matters. For example, objective tests are no longer used to screen astronauts, since they were not successful in predicting who would be successful or unsuccessful (Cordes, 1983). There does, in fact, appear to be a movement within the general public and the profession toward discontinuation of the use of personality assessment procedures for decision-making in employment situations. We would also note as another possibly significant trend a movement toward direct observation of behavior in the form of behavioral assessment, as in the case of the development of the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 1989). The zeitgeist definitely is in opposition to procedures in which the intent is disguised. Some time ago, Burdock and Zubin (1985) argued that, "nothing has as yet replaced behavior for evaluation of mental patients," and a similar argument might be made in current times.

## Neuropsychological assessment

Another area that has an interesting historical development is neuropsychological assessment. The term itself is a relatively new one and probably was made popular through the first edition of [Lezak's \(1976\)](#) book of that title. Neuropsychological assessment is of particular historical interest because it represents a confluence of two quite separate antecedents: central and eastern European behavioral neurology and American and English psychometrics. Neurologists, of course, have always been concerned with the behavioral manifestations of structural brain damage and the relationship between brain function and behavior. Broca's discovery of a speech center in the left frontal zone of the brain is often cited as the first scientific neuropsychological discovery because it delineated a relatively specific relationship between a behavioral function—that is, speech—and a correspondingly specific region of the brain (the third frontal convolution of the left hemisphere). Clinical psychologists developed an interest in this area when they were called upon to assess patients with known or suspected brain damage. The first approach to this diagnostic area involved utilization of the already-existing psychological tests, and the old literature deals primarily with how tests such as the Wechsler scales, the Rorschach test, or the Bender–Gestalt test could be used to diagnose brain damage. More recently, special tests were devised specifically for assessment work with patients having known or suspected brain damage. The merger between clinical psychology and behavioral neurology can be said to have occurred when the sophistication of neurologists working in the areas of brain function and brain disease was combined with the psychometric sophistication of clinical psychology. The wedding occurred when reliable, valid, and well-standardized measurement instruments began to be used to answer complex questions in neurological and differential neuropsychiatric diagnosis. Thus, clinicians who ultimately identified themselves as clinical neuropsychologists tended to be individuals who knew their psychometrics, but who also had extensive training and experience in neurological settings. Just as many clinical psychologists worked with psychiatrists, many clinical neuropsychologists worked with neurologists and neurosurgeons. This relationship culminated in the development of standard neuropsychological test batteries, notably the Halstead–Reitan ([Reitan & Wolfson, 1993](#)) and Luria–Nebraska batteries ([Golden, Hammeke, & Purisch, 1980](#); [Golden, Purisch, & Hammeke, 1985](#)), as well as in the capacity of many trained psychologists to perform individualized neuropsychological assessments of adults and children. Thus, within the history of psychological assessment, clinical neuropsychological evaluation has recently emerged as an independent discipline to be distinguished from general clinical psychology on the basis of the specific expertise that members of that discipline have in the areas of brain–behavior relationships and diseases of the nervous system. There have been expansions of both the standard batteries and the individual neuropsychological tests. An alternate form ([Golden, et al., 1985](#)) as well as a children's version ([Golden, 1981](#)) of the Luria–Nebraska Neuropsychological Battery are now available. Also prominent are the series of tests described in detail by Arthur Benton

and collaborators in Contributions to Neuropsychological Assessment (Benton, Hamsher, Vamey, & Spreen, 1983), the California Verbal Learning Test (Delis, Kramer, Kaplan, & Ober, 1987), and the recently revised and thoroughly reworked Wechsler Memory Scale (WMS-III and WMS-IV) (Wechsler, 1997a, 1997b, 2009). General distinctions have also developed between the approach to assessment, that is, whether one should use a fixed battery of tests that are administered to all patients regardless of diagnosis or referral question, or whether tests should be selected in an individualized manner based on the unique characteristics of the patient, or whether a balance should be struck between these two approaches so that a group of tests are administered to all patients to assess major domains of cognitive function with additional tests given based on the unique patient characteristics and referral question. Active discussions regarding the value or summative scores and actuarial approaches to test interpretation versus the process patients use to complete test items are also ongoing. Since the publication of the last edition of this handbook, many new neuropsychological tests have been developed, and two new comprehensive batteries have appeared, the Neuropsychological Assessment Battery (White & Stern, 2003) and the Meyers Neuropsychological system (Meyers & Rohling, 2004) that are now in common use. These procedures are described in the chapter on neuropsychological assessment batteries by Goldstein, Allen, and DeLuca.

## Behavioral assessment

Behavioral assessment has been one of the major developments to emerge in the field of psychological evaluation (Bellack & Hersen, 1988a, 1998b). Although its seeds were planted long before behavior therapy became a popular therapeutic movement, it is with the advent of behavior therapy that the strategies of behavioral assessment began to flourish (cf. Hersen & Bellack, 1976, 1981). Behavioral assessment can be conceptualized as a reaction to a number of factors (Barlow & Hersen, 1984; Hersen & Barlow, 1976; Hersen & Bellack, 1976). Among these were problems with unreliability and invalidity of aspects of the DSM-I and DSM-II diagnostic schemes and concerns over the indirect relationship between what was evaluated in traditional testing (e.g., the projective tests) and how it subsequently was used in treatment planning and application. Increasing acceptance of behavior therapy by the professional community as a viable series of therapeutic modalities, and parallel developments in the field of diagnosis in general, involving greater precision and accountability (e.g., the problem-oriented record (POR)) also fueled the movement toward behavioral assessment.

Among factors contributing to development of behavioral assessment, unreliability and invalidity of aspects of the DSM-I and DSM-II diagnostic schemes made early DSMs targets of considerable criticism from psychiatrists (Hines & Williams, 1975) and psychologists alike (Begelman, 1975). Indeed, Begelman (1975), in a more humorous vein, referred to the two systems as “twice-told tales” in the sense

that neither resulted in highly reliable classification schemes when patients were independently evaluated by separate psychiatric interviewers (cf. [Ash, 1949](#); [Sandifer, Pettus, & Quade, 1964](#)). Problems were especially evident when attempts to obtain interrater reliability were made for the more minor diagnostic groupings of the DSM schemes. Frequently, clinical psychologists would be consulted to carry out their testing procedures to confirm or disconfirm psychiatrists' diagnostic impressions based on DSM-I and DSM-II. But in so doing, such psychologists, operating very much as X-ray technicians, used procedures (objective and projective tests) that only had a tangential relationship to the psychiatric descriptors for each of the nosological groups of interest. Thus, over time, the futility of this kind of assessment strategy became increasingly apparent.

Moreover, not only were there problems with diagnostic reliability for the DSM-I and DSM-II, but empirical studies also documented considerable problems with regard to external validity of the systems ([Eisler & Polak, 1971](#); [Nathan, Zare, Simpson, & Ardberg, 1969](#)). Probably more important was the fact that the complicated psychological evaluation had a limited relationship to eventual treatment. At least in the psychiatric arena, the usual isomorphic relationship between assessment and treatment found in other branches of therapeutics did not seem to hold. The isolated and extended psychological examination was viewed by some as an empty academic exercise resulting in poetic jargon in the report that eventuated, whose utility was woefully limited as treatment seemed to be unrelated to the findings in the reports. These concerns resulted in attempts by clinical psychologists to measure the behaviors of interest in a direct fashion. For example, if a patient presented with a particular phobia, the objective of evaluation was not to assess the underlying "neurotic complex" or "alleged psychodynamics." Quite the contrary, the primary objective was to quantify in distance how closely the patient could approach the phobic object (i.e., the behavioral approach task) and how their heart rate (physiological assessment) increased as they got closer. In addition, the patient's cognitions (self-report) were quantified by having the patient assess their own level of fear (e.g., on a 1–10 point scale). Thus, the behavioral assessment triad, consisting of motoric, physiological, and self-report systems ([Hersen, 1973](#)), was established as the alternative to indirect measurement. [Hersen and Barlow \(1976\)](#) argue that whereas in indirect measurement a particular response is interpreted in terms of a presumed underlying disposition, a response obtained through direct measurement is simply viewed as a sample of a large population of similar responses elicited under those particular stimulus conditions. Thus, it is hardly surprising that proponents of direct measurement favor the observation of individuals in their natural surroundings whenever possible. When such naturalistic observations are not feasible, analogue situations approximating naturalistic conditions may be developed to study the behavior in question (e.g., the use of a behavioral avoidance test to study the degree of fear of snakes). When neither of these two methods is available or possible, subjects' self-reports are also used as independent criteria, although at times they may be operating under the control of totally different sets of contingencies than those governing motoric responses.

While the tripartite system of direct measurement is favored by behaviorists, it is in the realm of motoric behavior that behavior therapists have made the greatest and most innovative contributions (see [Foster, Bell-Dolan, & Burge, 1988](#); [Hersen, 1988](#); [Tryon, 1986](#)). With increased acceptance of behavior therapy, practitioners of the strategies found that their services were required in a large variety of educational, rehabilitation, community, medical, and psychiatric settings. Very often they were presented with extremely difficult educational, rehabilitation, and treatment cases, both from assessment and therapeutic perspectives. Many of the clients and patients requiring remediation exhibited behaviors that previously had not been measured in any direct fashion. Thus, there were few guidelines with regard to how the behavior might be observed, quantified, and coded. In many instances, "seat-of-the-pants" measurement systems were devised on the spot but with little regard for the psychometric qualities cherished by traditional assessment. If one peruses through the pages of the *Journal of Applied Behavior Analysis*, *Behaviour Research and Therapy*, *Journal of Behavior Therapy and Experimental Psychiatry*, and *Behavior Modification*, particularly in the earlier issues, numerous examples of innovative behavioral measures and more comprehensive systems are to be found. Consistent with the idiographic approach, many of these apply only to the case in question, have some internal or face validity, but, of course, have little generality or external validity. (Further comment on this aspect of behavioral assessment is made in a subsequent section of this chapter.)

A final development that contributed to and coincided with the emergence of behavioral assessment was the POR. This was a system of record-keeping first instituted on medical wards in general hospitals to sharpen and pinpoint diagnostic practices (cf. [Weed, 1964, 1968, 1969](#)). Later this system was transferred to psychiatric units (cf. [Hayes-Roth, Longabaugh, & Ryback, 1972](#); [Katz & Woolley, 1975](#); [Klonoff & Cox, 1975](#); [McLean & Miles, 1974](#); [Scales & Johnson, 1975](#)), with its relevance to behavioral assessment increasingly evident ([Atkinson, 1973](#); [Katz & Woolley, 1975](#)). When applied to psychiatry, the POR can be divided into four sections: (1) database, (2) problem list, (3) treatment plan, and (4) follow-up data. There can be no doubt that this kind of record-keeping promotes and enhances the relationship of assessment and treatment, essentially forcing the evaluator to crystallize their thinking about the diagnostic issues. In this regard, we have previously pointed out that despite the fact that POR represents, for psychiatry, a vast improvement over the type of record-keeping and diagnostic practice previously followed, the level of precision in describing problem behaviors and treatments to be used remedially does not yet approach the kind of precision reached in the carefully conducted behavioral analysis ([Hersen, 1976](#), p. 15). However, the POR certainly can be conceptualized as a major step in the right direction. In most psychiatric settings some type of POR (linking it to specific treatment plans) has been or is currently being used and, to a large extent, has further legitimized the tenets of behavioral assessment by clearly linking the problem list with specific treatment (cf. [Longabaugh, Fowler, Stout, & Kriebel, 1983](#); [Longabaugh, Stout, Kriebel, McCullough, & Bishop, 1986](#)).

## **Assessment schemes**

Since 1968 a number of comprehensive assessment schemes have been developed to facilitate the process of behavioral assessment (Cautela, 1968; Kanfer & Saslow, 1969; Lazarus, 1973). We outline a number of these to illustrate how the behavioral assessor conceptualizes his or her cases. Cautela (1968) depicted the role of behavioral assessment during three stages of treatment. In the first stage the clinician identifies adaptive behaviors and those antecedent conditions maintaining them. This step is accomplished through interviews, observation, and self-report questionnaires. The second stage involves selection of the appropriate treatment strategies, evaluation of their efficacy, and decision-making about when to terminate their application. In the third stage a meticulous follow-up of treatment outcome is recommended. This is done by examining motoric, physiological, and cognitive functioning of the client, in addition to independent confirmation of the client's progress by friends, relatives, and employers. A somewhat more complicated approach to initial evaluation was proposed by Kanfer and Saslow (1969), which involves seven steps. The first involves a determination as to whether a given behavior represents an excess, a deficit, or an asset. The second is a clarification of the problem and is based on the notion that in order to be maintained, maladjusted behavior requires continued support. The third is a motivational analysis in which reinforcing and aversive stimuli are identified. Fourth is a developmental analysis, focusing on biological, sociological, and behavioral changes. The fifth stage involves assessment of self-control and whether it can be used as a strategy during treatment. The sixth is an analysis of the client's interpersonal life, and the seventh is an evaluation of the patient's sociocultural—physical environment. In their initial scheme, Kanfer and Saslow (1969) viewed their system in complementary fashion to the existing diagnostic approach (i.e., DSM-II). They did not construe it as supplanting DSM-II, but did see their seven-part analysis as serving as a basis for arriving at decisions for precise behavioral interventions, thus producing a more isomorphic relationship between assessment and treatment. Subsequently, Kanfer and Grimm (1977) turned their attention to how the interview contributes to the overall behavioral assessment. In so doing, suggestions were made for organizing client complaints under five categories: (1) behavioral deficiencies, (2) behavioral excesses, (3) inappropriate environmental stimulus control, (4) inappropriate self-generated stimulus control, and (5) problematic reinforcement contingencies (p. 7). Lazarus (1973) proposed yet another behavioral assessment scheme with the somewhat humorous acronym of BASIC ID: B = behavior, A = affect, S = sensation, I = imagery, C = cognition, I = interpersonal relationship, and D = the need for pharmacological intervention (i.e., drugs) for some psychiatric patients. The major issue underscored by this diagnostic scheme is that if any of the elements were overlooked, assessment would be incomplete, thus resulting in only a partially effective treatment. To be fully comprehensive, deficits or surpluses for each of the categories need to be identified so that specific treatments can be targeted for each. This, then, should ensure the linear relationship between assessment and treatment, ostensibly absent from the nonbehavioral assessment schemes.

Despite development of these and other schemes (e.g., Bornstein, Bornstein, & Dawson, 1984), there is little in the way of their formal evaluation in empirical fashion. Although these schemes certainly appear to have a good bit of face validity, few studies have been devoted to evaluating concurrent and predictive validity. This, of course, is in contrast to the considerable effort to validate the third edition of DSM (i.e., DSM-III, 1980; Hersen & Turner, 1984) and its revisions (i.e., DSM-III-R, 1987; DSM-IV, 1994; DSM-5, 2013). In a somewhat different vein, Wolpe (1977) expressed concern about the manner in which behavioral assessment typically was being conducted, which he referred to as "The Achilles' Heel of Outcome Research in Behavior Therapy." He was especially concerned that too little attention had been devoted to evaluation of the antecedents of behaviors targeted for treatment, thus leading to a therapeutic approach that may be inappropriate. For example, in treating disorders such as depression (Wolpe, 1986) and phobia (Michelson, 1984, 1986) it seems obvious that each factor found operative in a particular patient needs to be treated by a program appropriate to it. Failure is predictable when intervention is exclusively based on an approach targeting only one antecedent factor, when other factors may be equally or more important in an individual case. To compare the effects of different treatments on assorted groupings of individuals with depression or phobia is about as informative as comparing the effects of different antibiotics on strep throat in the absence of a bacterial diagnosis. Blanket treatment that does not take into account antecedents undoubtedly should fail (Wolpe & Wright, 1988). But here too, the necessary research findings to document this are as yet forthcoming (see White, Turner, & Turkat, 1983).

Contrasted to the field of psychological assessment in general, behavioral assessment as a specialty has had a developmental history of about five decades. However, in these decades there have been some remarkable changes in the thinking of behavioral assessors. Probably as a strong overt reaction to the problems perceived by behavioral assessors in traditional psychological evaluation, many of the sound psychometric features of that tradition were initially abandoned. Indeed, in some instances it appears that "the baby was thrown out with the bath water." As we already have noted, consistent with the idiographic approach to evaluation and treatment, little concern was accorded to traditional issues of reliability and validity. (An exception was the obsessive concern with high interrater reliability of observations of motoric behavior.) This was particularly the case for the numerous self-report inventories that were developed early on to be consistent with the motoric targets of treatment (e.g., some of the fear survey schedules). There were many other aspects of traditional evaluation that also were given short shrift. Intelligence testing was eschewed, norms and developmental considerations were virtually ignored, and traditional psychiatric diagnosis was viewed as anathema to behavior therapy. However, since the late 1970s this "hard line" has been mollified. With publication of the second, third, and fourth editions of *Behavioral Assessment: A Practical Handbook* and the emergence of two assessment journals (*Behavioral Assessment* and *Journal of Psychopathology and Behavioral Assessment*), greater attention to psychometric principles has returned. An example includes evaluation of the external validity of role playing as an assessment strategy in the social skill

areas by Bellack and his colleagues (cf. [Bellack, Hersen, & Lamparski, 1979](#); [Bellack, Hersen, & Turner, 1979](#); [Bellack, Turner, Hersen, & Luber, 1980](#)). Also in numerous overviews the relevance of the psychometric tradition to behavioral assessment has been articulated with considerable vigor (e.g., [Adams & Turner, 1979](#); [Cone, 1977, 1988](#); [Haynes, 1978](#); [Nelson & Hayes, 1979](#); [Rosen, Sussman, Mueser, Lyons, & Davis, 1981](#)). Looking at behavioral assessment today from a historical perspective, it certainly appears as though the “baby” is being recovered from the discarded bath water. Also, there have been several calls for a broadened conceptualization of behavioral assessment (e.g., [Bellack & Hersen, 1998](#); [Hersen, 1988](#); [Hersen & Bellack, 1988](#); [Hersen & Last, 1989](#); [Hersen & Van Hassett, 1998](#)). Such broadening has been most noticeable with respect to the use of intelligence tests in behavioral assessment ([Nelson, 1980](#)), the relevance of neuropsychological evaluation for behavioral assessment ([Goldstein, 1979](#); [Horton, 1988](#)), the importance of developmental factors especially in child and adolescent behavioral assessment ([Edelbrock, 1984](#); [Harris & Ferrari, 1983](#); [Hersen & Last, 1989](#)), and the contribution that behavioral assessment can make to pinpointing psychiatric diagnosis ([Hersen, 1988](#); [Tryon, 1986, 1998](#)).

### ***DSMs and behavioral assessment***

In the earlier days of behavioral assessment, traditional psychiatric diagnosis was, for the most part, shunned. Behavioral assessors saw little relationship between what they were doing and the overall implicit goals of DSM-II. Moreover, as we have noted, categories subsumed under DSM-II had major problems with reliability and validity. So, consistent with cogent criticisms about the official diagnostic system, behavioral assessors tended to ignore it when possible. They continued to develop their strategies independently of DSM-II and the then-emerging DSM-III. In fact, some (e.g., [Adams, Doster, & Calhoun, 1977](#); [Cautela, 1973](#)) advocated totally new diagnostic formats altogether, but these never had a chance of being accepted by the general diagnostic community, given the political realities. In spite of its problems and limitations, with the emergence of DSM-III ([APA, 1980](#)), behavioral therapists and researchers appeared to have assumed a somewhat different posture (cf. [Hersen, 1988](#); [Hersen & Bellack, 1988](#); [Hersen & Turner, 1984](#); [Nelson, 1987](#)). Such positions have been articulated by a number of prominent behavior therapists, such as [Nathan \(1981\)](#) and [Kazdin \(1983\)](#). But the issues concerning DSM-III and behavioral assessment were most clearly summarized by [Taylor \(1983\)](#), a behavioral psychiatrist; “The new Diagnostic and Statistical Manual of the American Psychiatric Association” is a major improvement in psychiatric diagnosis over previous classification systems. Where symptomatic diagnoses are useful, as in relating an individual’s problem to the wealth of clinical and research data in abnormal psychology or in identifying conditions which require specific treatments, the DSM-III represented the best available system at the time. Many conceptual and practical problems remained with DSM-III; for instance, it retained a bias toward the medical model, included many conditions which should not fall into a psychiatric diagnostic system, and included descriptive axes that had

not been adequately validated and have subsequently been removed from the DSM-5. Nevertheless, behavior therapists are well advised to become familiar with and use the most recent DSM as part of behavioral assessment. We, of course, would argue that the same holds true for those who use the DSM system. We are fully in accord with Taylor's comments and believe that if behavior therapists wish to impact on the accepted nosological system, they are urged to work from within rather than from without. In this connection, Tryon (1986) has presented the field with a marvelous outline for how motoric measurements in both children and adults will enable the DSM categories to gain greater precision. He clearly shows how many of the diagnostic categories (e.g., depression; attention deficit hyperactivity disorders) have motoric referents that could be evaluated by behavioral assessors. However, much work of a normative nature (to determine lower and upper limits of normality) is still required before any impact on the DSM system will be felt (Tryon, 1989). We believe that such evaluation represents an enormous challenge to behavioral assessors that could result in a lasting contribution to the diagnostic arena.

## Summary

We have provided a brief historical overview of several major areas in psychological evaluation: intellectual, personality, neuropsychological, and behavioral assessment. Some of these areas have lengthy histories, and others are relatively young. However, it seems clear that the tools used by psychologists 50 years ago are generally different from those used now. Behavioral assessment techniques, structured psychiatric interviews, and standard, comprehensive neuropsychological test batteries are all relatively new. Furthermore, the computer has made significant inroads into the assessment field, with online testing, scoring, and interpretation a reality in many cases. Serious efforts have been made in recent years to link assessment more closely to treatment and other practical concerns. We may also note a trend toward direct acquisition of information and observation, as exemplified in many behavioral assessment techniques. Similarly, while neuropsychological assessment is still heavily dependent on the use of formal tests, there is increasing application of those tests in rehabilitation planning and in the association between neuropsychological test results and functional activities of daily living. We also note a corresponding decrease in interest in such matters as brain localization, particularly since the CT scan, MRI, and related brain-imaging procedures have solved much of that problem. We would prognosticate that psychological assessment will be increasingly concerned with automation, the direct observation of behavior, and the practical application of assessment results.

## Recent developments

At this writing in 2018 it has been almost a century and a half since Galton's whistle was described in 1876. The progress of psychological assessment has been

ongoing since that time and has undergone some dramatic changes. Its applications to various aspects of behavior are numerous with branches involving health, education, achievement and vocational functioning. Since the publication of the last edition of this handbook 19 years ago this progression has continued with some important new developments.

Various applications of modern technology have greatly altered how assessment is accomplished, particularly with regard to automation and computers. We now have assessment laboratories located in educational and clinical settings in which tests are self-administered and scored by computers and sometimes interpreted producing written reports. Tests may be taken in unusual environments, perhaps most notably on space shuttles. Testing may now be done on the internet such that a test taken by computer can be administered from some remote location ([Naglieri et al., 2004](#)). Since the appearance of the last edition of this Handbook, this automation and computerization have progressed at a rapid pace. There remains controversy about whether computer-generated interpretation should become a common practice, but there is little question that computers can now write elegant and extensive reports. Body monitoring can provide detailed information about behaviorally related physiological status.

Neuropsychological assessment has developed rapidly, and clinical neuropsychology has been designated by the American Psychological Association as a professional specialty, with its own Division, the Society for Clinical Neuropsychology (Division 40) and numerous books and journals. During the past several years the typical characteristics of the assessments done have changed, but there is a great deal of variability among clinicians. For theoretical and practical reasons the use of lengthy standard test batteries has diminished and briefer, more individualized assessments are more common. There have been changes in the focus of the assessment going from strong emphasis on identification of pathology to greater consideration of functional, behavioral matters. There appears to be a strong alliance between neuropsychology and neurodiagnostics largely involving neuroimaging. The introduction of the CT scan followed by MRI has greatly changed the nature of neuropsychological assessment.

In recent years there has been an important change in the field of academic achievement testing. The “No Child Left Behind” government policy appears to have elevated the significance of achievement test scores that are now used in essence to evaluate quality of educational programs. Associated controversies have arisen concerning the culture-fairness of these tests associated with development of a field of research aimed toward the creation of culture-fair tests. The chapters in this book by Puente and Melikyan and Katz and Brown deal with this area in detail. This matter remains an area of significant concern and the field is in something of a ferment. Similar controversy has arisen concerning the matter of the genetics of intelligence.

Personality assessment continues to lean toward objective tests and away from the projective techniques. The revised MMPI appears to have become the major instrument in common use, as pointed out in the chapter by Dr. Williams and colleagues whose title characterizes the MMPI-2 as “The predominant multidimensional

self-report inventory." An interest in structured interviewing continues and we anticipate appearance of new interviews that update their content to be consistent with the changes made in DSM-5. In the area of behavioral assessment motivational interviewing has become a method of great importance and is now widely used. While it is basically a counseling and not an assessment method, it utilizes questioning of the client about areas of desired change. Like much behavioral assessment it is directly associated with treatment.

It is apparent that there have been many recent changes in psychological assessment concerning several of its aspects. It is an evolving field, and the practices and methods of today may be different from those in use tomorrow. Its association with modern technology appears to be strengthening, and there appears to be increased concern with relation to treatment and adaptive functioning, particularly in clinical neuropsychological assessment. The process of revising assessment in the forms of new editions or developing new tests is continuing at a rapid pace.

## References

- Adams, H. E., Doster, J. A., & Calhoun, K. S. (1977). A psychologically-based system of response classification. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment*. New York: Wiley.
- Adams, H. E., & Turner, S. M. (1979). Editorial. *Journal of Behavioral Assessment*, 1, 1–2.
- Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt.
- American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.
- American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd Rev. ed.). Washington, DC: Author.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5)* (5th ed.). Washington, DC: Author.
- Andreasen, N. C. (1984). *Scale for the assessment of negative symptoms (SANS)*. Iowa City, IA: University of Iowa.
- Ash, P. (1949). The reliability of psychiatric diagnosis. *Journal of Abnormal and Social Psychology*, 44, 272–276.
- Atkinson, C. (1973). *Data collection and program evaluation using the problem-oriented medical record*. Miami, FL: Association for Advancement of Behavior Therapy.
- Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.
- Begelman, D. A. (1975). Ethical and legal issues in behavior modification. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 1). New York: Academic Press.
- Bellack, A. S., & Hersen, M. (1988a). Future directions. In A. S. Bellack, & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed.). New York: Pergamon Press.
- Bellack, A. S., & Hersen, M. (Eds.). (1998b). *Behavioral assessment: A practical handbook* (4th ed.). Needham Heights, MA: Allyn & Bacon.

- Bellack, A. S., Hersen, M., & Lamparski, D. (1979). Role-playing tests for assessing social skills: Are they valid? Are they useful? *Journal of Consulting and Clinical Psychology*, 47, 335, -.
- Bellack, A. S., Hersen, M., & Turner, S. M. (1979). Relationship of role playing and knowledge of appropriate behavior to assertion in the natural environment. *Journal of Consulting and Clinical Psychology*, 47, 679–685.
- Bellack, A. S., Turner, S. M., Hersen, M., & Luber, R. (1980). Effects of stress and retesting on role-playing tests of social skill. *Journal of Behavioral Assessment*, 2, 99–104.
- Benton, A. L., Hamsher, K. de S., Vamey, N. R., & Spreen, O. (1983). *Contributions to neuropsychological assessment: A clinical manual*. New York: Oxford University Press.
- Blessed, G., Tomlinson, B. E., & Roth, M. (1968). The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry*, 114, 797–811.
- Boring, E. G. (1950). *A history of experimental psychology*. New York: Appleton-Century-Crofts.
- Bornstein, P. H., Bornstein, M. T., & Dawson, B. (1984). Integrated assessment and treatment. In T. H. Ollendick, & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures*. New York: Pergamon Press.
- Burdock, E. I., Hardesty, A. S., Hakerem, G., Zubin, J., & Beck, Y. M. (1968). *Ward behavior inventory*. New York: Springer.
- Burdock, E. I., & Zubin, J. (1985). *Objective evaluation in psychiatry. Psychiatric reference and record book* (2nd ed). New York: Roerig Laboratories, Inc.
- Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the sixteen personality factor questionnaire. (Technical Report)*. Champaign, IL: Institute for Personality and Ability Testing.
- Cautela, J. R. (1968). Behavior therapy and the need for behavior assessment. *Psychotherapy: Theory, Research and Practice*, 5, 175–179.
- Cautela, J. R. (1973). A behavioral coding system. Presidential address presented at the seventh annual meeting of the Association for Advancement of Behavioral Therapy, Miami, FL.
- Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavioral Therapy*, 8, 411–426.
- Cone, J. D. (1988). Psychometric considerations and the multiple models of behavioral assessment. In A. S. Bellack, & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed). New York: Pergamon Press.
- Cordes, C. (1983). Mullane: Tests are grounded. *APA Monitor*, 14, 24.
- Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed). New York: Harper & Brothers, Original work published 1949.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *CVLT: California verbal learning test: Research Edition [Manual]*. San Antonio, TX: The Psychological Corporation.
- Edelbrock, C. (1984). Diagnostic issues. In T. H. Ollendick, & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 30–37). New York: Pergamon Press.
- Eisler, R. M., & Polak, P. R. (1971). Social stress and psychiatric disorder. *Journal of Nervous and Mental Disease*, 153, 227–233.
- Feighner, J., Robins, E., Guze, S., Woodruff, R., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry*, 26, 57–63.

- 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.
- Foster, S. L., Bell-Dolan, D. J., & Burge, D. A. (1988). Behavioral observation. In A. S. Bellack, & M. Hersen (Eds.), *Behavioral assessment: A practical handbook*. New York: Pergamon Press.
- Gilberstadt, H., & Duker, J. (1965). *A handbook for clinical and actuarial MMPI interpretation*. Philadelphia, PA: Saunders.
- Golden, C. J., Hammeke, T. A., & Purisch, A. D. (1980). *The Luria-Nebraska battery manual*. Los Angeles, CA: Western Psychological Services.
- Golden, C. J., Purisch, A. D., & Hammeke, T. A. (1985). *The Luria-Nebraska battery: Forms I and II*. Los Angeles, CA: Western Psychological Services.
- Golden, G. (1981). The Luria-Nebraska children's battery: Theory and formulation. In G. W. Hynd, & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-aged child: Issues and procedures*. New York: Grune & Stratton.
- Goldstein, G. (1979). Methodological and theoretical issues in neuropsychological assessment. *Journal of Behavioral Assessment, 1*, 23–41.
- Guilford, J. P., & Zimmerman, W. (1949). *Guilford-Zimmerman temperament survey*. Los Angeles, CA: Western Psychological Services.
- Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry, 23*, 56–62.
- Harris, S. L., & Ferrari, M. (1983). Developmental factors in child behavior therapy. *Behavior Therapy, 14*, 54–72.
- Hayes-Roth, F., Longabaugh, R., & Ryback, R. (1972). The problem-oriented medical record and psychiatry. *British Journal of Psychiatry, 121*, 27–34.
- Haynes, S. N. (1978). *Principles of behavioral assessment*. New York: Gardner Press.
- Helzer, J., Robins, L., Croughan, J., & Welner, A. (1981). Renard diagnostic interview. *Archives of General Psychiatry, 38*, 393–398.
- Hersen, M. (1973). Self-assessment and fear. *Behavior Therapy, 4*, 241–257.
- Hersen, M. (1976). Historical perspectives in behavioral assessment. In M. Hersen, & A. S. Bellack (Eds.), *Behavioral assessment designs: Strategies for studying behavior change*. New York: Pergamon Press.
- Hersen, M. (1988). Behavioral assessment and psychiatric diagnosis. *Behavioral Assessment, 10*, 107–121.
- Hersen, M., & Barlow, D. H. (1976). *Single-case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press.
- Hersen, M., & Bellack, A. S. (Eds.). (1976). *Behavioral assessment: A practical handbook* (1st ed.). New York: Pergamon Press.
- Hersen, M., & Bellack, A. S. (1981). *Behavioral assessment: A practical handbook* (2nd ed.). New York: Pergamon Press.
- Hersen, M., & Bellack, A. S. (1988). DSM-III and behavioral assessment. In A. S. Bellack, & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed.). New York: Pergamon Press.
- Hersen, M., & Last, C. G. (1989). Psychiatric diagnosis and behavioral assessment in children. In M. Hersen, & C. G. Last (Eds.), *Handbook of child psychiatric diagnosis*. New York: John Wiley & Sons.
- Hersen, M., & Turner, S. M. (1984). DSM-III and behavior therapy. In S. M. Turner, & M. Hersen (Eds.), *Adult psychopathology: A behavioral perspective*. New York: Wiley.

- Hersen, M., & Van Hassett, V. (Eds.). (1998). *Basic interviewing: A practical guide for counselors and clinicians*. Mahwah, NJ: Erlbaum.
- Hines, F. R., & Williams, R. B. (1975). Dimensional diagnosis and the medical students' grasp of psychiatry. *Archives of General Psychiatry*, 32, 525–528.
- Holtzman, W. H. (1958). *The Holtzman Inkblot technique*. New York: Psychological Corporation.
- Honigfeld, G., & Klett, C. (1965). The Nurse's Observation Scale for Impatient Evaluation (NOSIE): A new scale for measuring improvement in schizophrenia. *Journal of Clinical Psychology*, 21, 65–71.
- Horton, A. M. (1988). Use of neuropsychological testing in determining effectiveness of ritalin therapy in an DDRT patient. *Behavior Therapist*, 11, 114–118.
- Kanfer, F. H., & Grimm, L. G. (1977). Behavior analysis: Selecting target behaviors in the interview. *Behavior Modification*, 1, 7–28.
- Kanfer, F. H., & Saslow, G. (1969). Behavioral diagnosis. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill.
- Katz, R. C., & Woolley, F. R. (1975). Improving patients' records through problem orientation. *Behavior Therapy*, 6, 119–124.
- Kazdin, A. E. (1983). Psychiatric diagnosis, dimensions of dysfunction, and child behavior therapy. *Behavior Therapy*, 14, 73–99.
- Kirkpatrick, B., Strauss, G. P., Nguyen, L., Fischer, B. A., Daniel, D. G., Cienfuegos, A., & Marder, S. R. (2011). The brief negative symptom scale: Psychometric properties. *Schizophrenia Bulletin*, 37, 300–305.
- Klonoff, H., & Cox, B. (1975). A problem-oriented system approach to analysis of treatment outcome. *American Journal of Psychiatry*, 132, 841–846.
- Kring, A. M., Gur, R. E., Blanchard, J. J., Horan, W. P., & Reise, S. P. (2013). The Clinical Assessment Interview for Negative Symptoms (CAINS): Final development and validation. *American Journal of Psychiatry*, 170(2), 165–172.
- Lazarus, A. A. (1973). Multimodal behavior therapy: Treating the "basic id". *Journal of Nervous and Mental Disease*, 156, 404–411.
- Lezak, M. (1976). *Neuropsychological assessment*. New York: Oxford University Press.
- Longabaugh, R., Fowler, D. R., Stout, R., & Kriebel, G. (1983). Validation of a problem-focused nomenclature. *Archives of General Psychiatry* (40, pp. 453–461).
- Lindzey, G. (1965). Seer vs sign. *Journal of Experimental Research in Personality*, 1, 17–26.
- Longabaugh, R., Stout, R., Kriebel, G. M., McCullough, L., & Bishop, D. (1986). DSM-III and clinically identified problems as a guide to treatment. *Archives of General Psychiatry*, 43, 1097–1103.
- Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19, 185–212.
- Machover, K. (1949). *Personality projection in the drawing of the human figure: A method of personality investigation*. Springfield, IL: Charles Thomas.
- Marks, P. A., Seeman, W., & Hailer, D. L. (1974). *The actuarial use of the MMPI with adolescents and adults*. Baltimore, MD: Williams & Wilkins.
- McFie, J. (1975). *Assessment of organic intellectual impairment*. London: Academic Press.
- McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. (2012, pp. 151–179). New York: Guilford Press.

- McLean, P. D., & Miles, J. E. (1974). Evaluation and the problem-oriented record in psychiatry. *Archives of General Psychiatry*, 31, 622–625.
- Meehl, P. E. (1954). *Clinical vs. statistical prediction*. Minneapolis, MN: University of Minnesota Press.
- Menninger, K. A. (1952). *A manual for psychiatric case study*. New York: Grune & Stratton.
- Meyers, J. E., & Rohling, M. L. (2004). Validation of the Meyers short battery on mild TBI patients. *Archives of Clinical Neuropsychology*, 19, 637–651.
- Michelson, L. (1984). The role of individual differences, response profiles, and treatment consonance in anxiety disorders. *Journal of Behavioral Assessment*, 6, 349–367.
- Michelson, L. (1986). Treatment consonance and response profiles in agoraphobia: Behavioral and physiological treatments. *Behaviour Research and Therapy*, 24, 263–275.
- Mihura, J. L., Meyer, G. J., Dumitrascu, N., & Bombel, G. (2013). The validity of individual Rorschach variables: systematic reviews and meta-analyses of the comprehensive system. *Psychological Bulletin*, 139(3), 548–605. Available from <https://doi.org/10.1037/a0029406>.
- Millon, T. (1982). *Millon clinical multiaxial inventory* (3rd ed.). Minneapolis, MN: National Computer Systems.
- Millon, T. (1985). The MCMI provides a good assessment of DSM-III disorders: The MCMI-II will prove even better. *Journal of Personality Assessment*, 49, 379–391.
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.
- Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the internet: New problems, old issues. *American Psychologist*, 59, 150–162.
- Nathan, P. E. (1981). Symptomatic diagnosis and behavioral assessment: A synthesis. In D. H. Barlow (Ed.), *Behavioral assessment of adult disorders*. New York: Guilford.
- Nathan, P. E., Zare, N. C., Simpson, H. F., & Ardberg, M. M. (1969). A systems analytic model of diagnosis: I. The diagnostic validity of abnormal psychomotor behavior. *Journal of Clinical Psychology*, 25, 3–9.
- Nelson, R. O. (1980). The use of intelligence tests within behavioral assessment. *Behavioral Assessment*, 2, 417–423.
- Nelson, R. O. (1987). DSM-III and behavioral assessment. In C. G. Last, & M. Hersen (Eds.), *Issues in diagnostic research*. New York: Plenum Press.
- Nelson, R. O., & Hayes, S. C. (1979). Some current dimensions on behavioral assessment. *Behavioral Assessment*, 1, 1–16.
- Overall, J. E., & Gorham, J. R. (1962). The brief psychiatric rating scale. *Psychological Reports*, 10, 799–812.
- Ozer, S., Young, J., Champ, C., & Burke, M. (2016). A systematic review of the diagnostic test accuracy of brief cognitive tests to detect amnestic mild cognitive impairment. *International Journal of Geriatric Psychiatry*, February 18. Available from <https://doi.org/10.1002/gps.4444>, [Epub ahead of print].
- Rapaport, D., Gill, M., & Schafer, R. (1945). *The Menninger Clinic monograph series. Diagnostic psychological testing*, Vol. 1. Chicago, IL, US: Year Book Publishers.
- Raskin, A. (1982). Assessment of psychopathology by the nurse or psychiatric aide. In E. I. Burdock, A. Sudilovsky, & S. Gershon (Eds.), *The behavior of psychiatric patients: Quantitative techniques for evaluation*. New York: Marcel Dekker.
- Reitan, R. M., & Wolfson, D. (1993). *The Halstead–Reitan neuropsychological test battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

- Rohde, A. R. (1957). *The sentence completion method*. New York: Ronald Press.
- Rollnick, S., & Miller, W. R. (1995). Motivational interviewing. *Behavioural and Cognitive Psychotherapy*, 23, 325–334.
- Rosen, A. J., Sussman, S., Mueser, K. T., Lyons, J. S., & Davis, J. M. (1981). Behavioral assessment of psychiatric inpatients and normal controls across different environmental contexts. *Journal of Behavioral Assessment*, 3, 25–36.
- Sandifer, M. G., Jr., Pettus, C., & Quade, D. (1964). A study of psychiatric diagnosis. *Journal of Nervous and Mental Disease*, 139, 350–356.
- Scales, E. J., & Johnson, M. S. (1975). A psychiatric POMR for use by a multidisciplinary team. *Hospital and Community Psychiatry*, 26, 371–373.
- Shneidman, E. S. (1952). *Make a picture test*. New York: The Psychological Corporation.
- Spitzer, R. L., & Endicott, J. (1977). *Schedule for affective disorders and schizophrenia (Technical Report)*. New York: New York State Psychiatric Institute, Biometrics Research Department.
- Spitzer, R. L., Endicott, J., & Robins, E. (1977). *Research diagnostic criteria (RDC) for a selected group of functional disorders*. Bethesda, MD: National Institute of Mental Health.
- Spitzer, R. L., & Williams, J. B. W. (1983). *Instruction manual for the structured clinical interview for DSM-III (SCID)*. New York: New York State Psychiatric Institute, Biometrics Research Department.
- Swensen, C. H. (1957). Empirical evaluations of human figure drawings, 1957–1966. *Psychological Bulletin*, 54, 431–466.
- Swensen, C. H. (1968). Empirical evaluations of human figure drawings. *Psychological Bulletin*, 20, 20–44.
- Szondi, L. (1952). *Experimental diagnostics of drives*. New York: Grune & Stratton.
- Taylor, C. B. (1983). *DSM-III and behavioral assessment*. *Behavioral Assessment* (5, pp. 5–14).
- Tryon, W. W. (1986). Motor activity measurements and DSM-III. In M. Hersen (Ed.), *Innovations in child behavior therapy*. New York: Springer.
- Tryon, W. W. (1989). Behavioral assessment and psychiatric diagnosis. In M. Hersen (Ed.), *Innovations in child behavior therapy*. New York: Springer.
- Tryon, W. W. (1998). In A. S. Bellack, & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Needham Heights, MA: Allyn & Bacon.
- VanLnenep, D. J. (1951). The four-picture test. In H. H. Anderson, & G. L. Anderson (Eds.), *An introduction to projective techniques*. New York: Prentice-Hall.
- Wechsler, D. (1944). *The measurement of adult intelligence*. Baltimore, MD: Williams & Wilkins.
- Wechsler, D. (1997a). *WAIS-III administration and scoring manual*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997b). *Wechsler memory scale III (WMS-III)*. San Antonio, TX: The Psychological Corporation.
- Weed, L. L. (1964). Medical records, patient care, and medical education. *Irish Journal of Medical Sciences*, 6, 271–282.
- Weed, L. L. (1968). Medical records that guide and teach. *New England Journal of Medicine*, 278, 593–600.
- Weed, L. L. (1969). *Medical records, medical education, and patient care*. Cleveland, OH: Case Western Reserve University Press.
- White, T., & Stern, R. A. (2003). *NAB: Neuropsychological assessment battery*. Lutz, FL: Psychological Assessment Resources.

- White, D. K., Turner, L. B., & Turkat, I. D. (1983). The etiology of behavior: Survey data on behavior therapists' contributions. *The Behavior Therapist*, 6, 59–60.
- Wolpe, J. (1977). Inadequate behavior analysis: The Achilles heel of outcome research in behavior therapy. *Journal of Behavior Therapy and Experimental Psychiatry*, 8, 1–3.
- Wolpe, J. (1986). The positive diagnosis of neurotic depression as an etiological category. *Comprehensive Psychiatry*, 27, 449–460.
- Wolpe, J., & Wright, R. (1988). The neglect of data gathering instruments in behavior therapy practice. *Journal of Behavior Therapy and Experimental Psychiatry*, 19, 5–9.
- Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: reliability, validity and sensitivity. *British Journal of Psychiatry*, 133(5), 429–435.
- Zubin, J. (1967). *Classification of the behavior disorders*. Annual Review of Psychology (18, pp. 373–406). Palo Alto, CA: Annual Review, Inc.
- Zubin, J. (1984). Inkblots do not a test make. *Contemporary Psychology*, 29, 153–154.

## Further reading

- Bemhardt, A. J., Hersen, M., & Barlow, D. H. (1972). Measurement and modification of spasmodic torticollis: An experiment analysis. *Behavior Therapy*, 3, 294–297.
- Nelson, R. O. (1979). DSM-III and behavioral assessment. In M. Hersen, & C. G. Last (Eds.), *Issues in diagnostic research*. New York: Plenum Press.
- Robins, S. L., Helzer, J., Croughan, N. A., & Ratcliff, K. (1981). National institute of mental health diagnostic interview schedule. *Archives of General Psychiatry*, 38, 381–389.

## **Part II**

# **Psychometric Foundations**

# How to develop an empirically based psychological test

2

Cecil Reynolds<sup>1</sup> and Ron Livingston<sup>2</sup>

<sup>1</sup>Texas A&M University, Austin, TX, United States, <sup>2</sup>The University of Texas at Tyler, Tyler, TX, United States

## Introduction

In this chapter we present a general model of test development that can be applied to most types of tests. The chapter emphasizes the development of a strategic plan for test development rather than a step-by-step description of implementing the plan. We have found that plan implementation varies greatly depending on the setting, the type of test being developed, and the level of support available. However, our experience is that a detailed, comprehensive plan is necessary for—and typically results in—successful implementation. To build anything of substance you must start with a good plan!

The validity of test score interpretations is the most important consideration in test development, and so ensuring the validity of those interpretations must begin when you initiate the process of test development. The approach to test development presented in this chapter is based on the experiences of the authors in developing commercial tests and from working with numerous test publishers. We focus on how commercial tests are designed and implemented, since they are the most common tests encountered beyond teacher-made classroom tests. However, this model may be implemented on a smaller scale for those developing a test for smaller research projects.

In this chapter we describe four broad stages—or phases—for developing tests. These are: Test Conceptualization; Specification of Format and Structure; Specification of Standardization, Scaling, and Psychometric Studies; and Plan Implementation. The first phase, Test Conceptualization, is designed to give you and your development team (test development is almost always a team effort) a clear understanding of the nature of the test, the construct it will measure, and how and why it will be used. This information drives many of the subsequent aspects of the test design. [Table 2.1](#) provides an outline for this first phase of the test development process.

**Table 2.1** Phase I—test conceptualization

Step 1: Specify the construct and establish a need
Step 2: Specify proposed applications
Step 3: Specify users
Step 4: Specify conceptual and operational definitions of construct

## Phase I: Test conceptualization

### ***Establish a need***

The first step in developing a psychological test is to establish the need for a new test. This includes specifying the construct (or constructs) you want to measure and establishing that there is actually a need for a new test to measure it. At this point a general idea of the construct you intend to measure is sufficient, but in a subsequent step you will need to develop a more precise definition. Check the *Mental Measurements Yearbook* and you will see there are literally thousands of commercial tests available. There are also a large number of tests published in professional journals and on the internet—probably even more than the number of commercial tests—that measure narrow constructs or are designed for highly specific applications. A thorough review of the existing measures is necessary to determine if a new measure is actually needed. Your review of the literature might reveal that there are existing instruments that will serve your intended purpose without exacting the time, cost, and effort of developing a new measure. However, your review might reveal that there are no existing measures that fill the niche you are interested in, or that while there are existing measures, they have inadequate or outdated normative data (e.g., consider the Flynn Effect), unacceptable psychometric properties (e.g., reliability and validity), inadequate documentation, etc., and a new measure is clearly warranted.

There are many factors that influence the need for new tests. As psychology develops as a science constructs evolve and are modified. Consider how the construct of intelligence has evolved over the last 100 plus years. In the late 1800s intelligence was measured with simple measures of sensory and motor abilities. Over the 20th century different theories and models of intelligence emerged and new intelligence tests were developed to match the evolving construct. Each era saw the development and spread of new intelligence tests, with the tests changing as our thoughts about intelligence changed. Some intelligence tests were revised to match the new models while others could not be adjusted to accommodate current models and became obsolete. At the same time, new tests of intelligence were developed that matched modern theories of intelligence. This is clearly demonstrated with the advent of the Catell–Horn–Carroll (CHC) Theory of Intelligence which is a prominent contemporary model that has influenced the development of many modern intelligence tests. This same process has occurred in other areas of psychological assessment. For example, when the Minnesota Multiphasic Personality Inventory ([Hathaway & McKinley, 1940, 1943](#)) was originally

developed it was based on a taxonomy of psychopathology that included descriptors such as psychasthenia and hysteria, terms that have long since fallen out of common use. As the curricula in schools change, tests of academic achievement must change; similarly, as new models of personality emerge (e.g., Five-Factor Model of Personality) new measures of personality are needed, and so on. In addition to evolving traditional constructs, new psychological constructs emerge. These new constructs are usually derived from theory and observation and must be empirically studied to be understood. To be studied, they must be measured, and to be measured we need tests.

Advances in technology and test theory may allow us to develop better or more precise methods of measuring a construct. Advances in technology may permit testing formats that were not previously available allowing test developers to devise improved techniques for measuring constructs. For example, reaction time was once measured through human observation and the use of a stop watch. This is certainly rudimentary when compared to the computer controlled presentation of stimuli and precise electronic measurement in milliseconds that is currently available. Accordingly, advances in test theory (e.g., Item Response Theory; generalizability theory) may allow the development of new assessment techniques.

In summary, there is always a need for new tests, whether it is to improve and modernize existing instruments or to develop new instruments to measure new constructs. In addressing the question of need, [Reynolds and Livingston \(2012\)](#) suggest asking: (1) Will the test improve psychological practice or research? and, (2) Will it improve the human condition?

### ***Specify proposed applications***

After defining the construct and establishing the need for a new test, you should describe how the test will be used and how the results will be interpreted. For example, will the test scores be used to predict performance on a criterion, facilitate clinical diagnosis, enhance treatment planning, improve employee selection, etc.? In what settings will the test be used? Will the primary setting be clinical, school, forensic, industrial—organizational, etc.? Naturally the answers to these questions should flow from the previous step and they will provide important guidance for subsequent steps in developing the test.

Consider the development of an omnibus personality scale. Will the proposed test focus on normal personality development or on psychopathological states? Different applications and settings necessitate different content as well as different interpretive schemes. At times even legal statutes directly influence test development. For example, tests designed to assist in personnel selection might appear similar to tests designed to facilitate clinical diagnosis; however, during development tests used in personnel selection must be carefully scrutinized for items that are illegal to ask in preemployment screening. Knowing the application, setting, and purpose of a test will directly impact many of the subsequent steps outlined in this chapter, from who will use the test to the appropriate normative sample to the types of validity studies that are needed.

## ***Specify users***

It is important to design tests with specific examiners in mind and the test manual should include a section that specifies the expected qualifications of the examiner. In most states there are licensure and certification requirements that restrict the use of psychological tests to specific professionals such as psychologists or licensed professional counselors. However, these requirements vary from state to state, and to deal with this variability we recommend that test authors focus more on the expected formal academic training and supervised experience of the test user as opposed to specifying licenses, certificates, and/or job titles.

Specifying the user of the test will also guide many subsequent decisions about test features such as the item format, administration, scoring, and interpretation. Different classes of users have different levels of training and skills and the test must be matched to the profile of the intended users. For example, a test designed to detect the presence of a clinical diagnosis would likely be targeted for clinical psychologists, while a test designed to screen a large number of students for reading problems might be intended for classroom teachers. Knowing the anticipated user allows the author to tailor the test to that user's specific profile and set of skills.

## ***Specify conceptual and operational definitions of constructs***

We often believe we understand the meaning of constructs like anxiety, depression, intelligence, hyperactivity, etc., until we try to put them into words and then realize we do not have them as clearly defined as we originally thought. Developing clear definitions of the construct(s) we want to measure helps us to clarify to ourselves what we want to measure and also allows us to explain it more clearly to others. We also refer to these definitions throughout the test development process to guide us in writing and selecting items, and later in interpreting the scores. We recommend developing two types of definitions—a conceptual and an operational definition. As the names suggest, the conceptual definition describes our construct at a theoretical level and the operational definition describes more specifically how our test will measure the construct. To illustrate this point we will provide an example using the common psychological concept of anxiety.

1. *Conceptual definition:* Anxiety is a psychological state that involves excessive worry and apprehension about a wide range of events and activities.
2. *Operational definition:* On the Anxiety Scale, anxiety will be measured by summing ratings in the specified direction on items that reflect worry, fear, apprehension, muscle tension, difficulty relaxing, restlessness, irritability, difficulty concentrating, and sleep disturbance.

The conceptual definition describes what we want to measure in the abstract while the operational definition provides more specific information about how the construct will be defined and measured in our test.

**Table 2.2** Phase II—specify test format and structure

- |   |
|---|
| Step 1: Specify age range   |
| Step 2: Specify test format   |
| Step 3: Specify internal structure (including scales of dissimulation)        |
| Step 4: Develop table of specification  |
| Step 5: Specify item format   |
| Step 6: Estimate how many items to include                                    |
| Step 7: Specify methods for item development, tryout, analysis, and selection |

## Phase II: Specify test format and structure

Once you have clearly specified the construct to be measured and the applications of your test, the next step is to delineate the structure and format of the test. This is where you identify the intended examinees (e.g., age ranges), the general format of the test (e.g., group vs individual administration), and the general structure of the test. This is also the appropriate time to develop a detailed test blueprint or table of specifications, specify the type of items to be included and how they will be developed and evaluated, and outline instructions for administering and scoring the test. [Table 2.2](#) provides an outline of the steps involved with this phase.

### ***Specify age range***

It is necessary to identify the appropriate age range for the test early in your test description. This factor will impact other features of the test such as the overall format of the test, its length, and the type of items you include. It will also guide decisions about the standardization sample you will need. For example, young children may have limited reading skills and not be able to respond accurately to self-report typical-response items (e.g., anxiety scale). Accordingly, unless you are measuring motor speed or hand–eye coordination, including items that require these abilities may compromise performance of elderly examinees. Overall test length may also be impacted by age. For example, young or elderly examinees might not have the attention span and stamina required to complete longer tests, so unless you are measuring those constructs, limited test length might be an important consideration. In summary, it is important to clearly identify the age range your test is intended for as it impacts many features of the test.

### ***Specify test format***

In this context the format of the test typically involves a number of different features. For example, will the test be administered individually or can it be administered to groups? This often depends on how much interaction is required between the examiner and the examinee. If extensive interaction is required an individually administered test is indicated, but if interactions are not necessary then group administration might be appropriate, and group administrations typically provide

more flexibility. Similarly, is the test to be presented orally, in print, or presented on a computer, iPad or similar device? You also want to specify who will be actually completing the test or protocol: the examiner, the examinee, or a third party informant?

### ***Specify internal structure***

Specifying the structure of the test involves indicating what scores you expect the test to produce. Will the test yield just one score that is a composite of all of the items or will there be multiple subscales? If there are multiple subscales, then provide specific information on what each one measures, and their relationship to each other. Naturally this is directly linked to the construct(s) you are measuring and the way you originally defined it (them) in Phase I. It is important to recognize that at this point you are specifying a hypothesized structure and this structure might change as empirical analyses provide information on the actual internal structure of the test. This is also the time to specify the techniques that will be used to examine the internal structure of the test. For example, will you be using exploratory and/or confirmatory factor analysis to examine the internal structure of the test?

This is also the appropriate time to decide if you will include scales for detecting dissimulation in your test. Dissimulation is the presentation of one's self in an inaccurate manner. For example, if I am really feeling anxious but do not want to admit it, I might respond to an item such as "I feel anxious" as false to conceal my true feelings. It is also possible for informants to engage in dissimulation of others if they are completing a rating scale about another person such as their child or spouse. There are many reasons people engage in dissimulation. People might deny symptoms on a personality test because they do not want to admit to feelings and behaviors others might consider unacceptable. In employment settings job applicants often attempt to respond in the manner they believe will most likely get them a job. Examinees will also endorse the presence of symptoms they do not have in order to look more pathological or impaired than they actually are. This also can occur for many reasons and is termed malingering when they have something to gain by making a false presentation. For example, an examinee may fake psychological problems and cognitive deficits in order to obtain disability benefits, increase damage rewards in civil case, or avoid or reduce punishment for crimes they committed.

With typical-response tests (e.g., personality tests) dissimulation scales are also referred to as validity scales, with the most common being F-Scales, L-Scales, and inconsistency scales. These are briefly described below:

1. F-scales are also referred to as "infrequency" scales because they contain items that are rarely endorsed, even by individuals displaying significant psychopathology. These scales are designed to have a low average intercorrelation among the items, indicating that they reflect disparate symptoms, not a coherent pattern. As a result, when an examinee endorses a large number of items it is likely they are exaggerating the presence of various unrelated symptoms in an attempt to appear pathological (i.e., faking bad). On F-scales, like other scales designed to detect dissimulation, it is cumulative response pattern that is

examined, not the response to a single item. However, elevations on an F-scale are not synonymous with malingering as many are wont to ascribe, and require careful scrutiny for accurate interpretation. Elevations may be a “plea for help” in a person in acute distress, or may reflect accurately the severity of a person’s condition. A detailed history and corroborative concurrent data are necessary to derive the correct interpretation.

2. L-scales were once referred to as “Lie Scales” but now are referred to appropriately as “Social Desirability” or “Fake Good” scales. These scales are designed to detect a response bias where the examinee attempts to deny or minimize the presence of pathology or undesirable traits. A social desirability scale might contain items written specifically for the scale such as “The judicial system never makes mistakes.” Most examinees responding in an honest manner will not agree with this, but an examinee attempting to appear particularly compliant might indicate agreement. These scales might also contain conventional items assessing feelings and behaviors most individuals are willing to acknowledge. For example, on the item “I get frustrated with other people” an examinee trying to appear socially appropriate might respond “Never,” but few individuals responding honestly would indicate they never get frustrated. Instead, most would acknowledge they get frustrated occasionally (if not more often).
3. Inconsistency scales are designed to detect inconsistent response patterns. Inconsistent responding might be the result of dissimulation but it also might be the result of other factors such as difficulty reading or comprehending the test items. Inconsistency scales are typically developed by identifying items that demonstrate high correlations and then comparing performance on them. As with the other dissimulation scales, it is the cumulative response pattern that is evaluated, not inconsistent responding on a single item pair.

Dissimulation scales are not as common in maximum-performance tests as they are in typical-response tests. A maximum-performance test is one, such as an IQ or achievement measure or any neuropsychological test, where the objective is to obtain the best possible or maximum level of performance from an examinee. A typical performance test asks the examinee to respond however they would usually or typically respond to the questions presented. In cognitive assessments the evaluation of dissimulation is often accomplished through effort testing. Effort testing involves giving tests that are easily accomplished by nearly any examinee if they attempt to complete the task accurately. Effort testing is recommended in most forensic assessments and whenever a patient has something to gain by feigning cognitive deficits ([Bush et al., 2005](#)). This has been a brief overview of scales for detecting dissimulation and a more detailed overview is provided by [Reynolds and Livingston \(2012\)](#).

### ***Develop a table of specifications or test blueprint***

The next step in test development is creating a Table of Specifications (TOSs) or Test Blueprint. The TOS helps ensure congruence between the construct(s) we intend to measure and the content of the test. TOSs were originally developed for achievement tests to ensure congruence between the curriculum and the test content. An example is given in [Table 2.3](#) for a hypothetical chapter on reliability. The column on the left, labeled content area, lists the major areas to be covered in the test. Across the top of the table we list the levels of Bloom’s cognitive taxonomy

**Table 2.3** Sample table of specifications for test on chapter on reliability

Content areas	Level of objective						Total
	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	
Measurement error	3	3		1	1		8
Classical test theory		2		2	2		6
Types of reliability estimates	2	2	4	2	2		12
Standard error of measurement	2	2	2	2	1		9

([Bloom, Englehart, Furst, Hill, & Krathwohl, 1956](#)). Bloom's taxonomy includes six processes in this taxonomy: knowledge, comprehension, application, analysis, synthesis, and evaluation. Knowledge is the most basic process, with each subsequent process being more advanced. The inclusion of these columns encourages you to consider the complexity of the processes you want to measure. Test developers tend to overemphasize lower level cognitive processes and to underemphasize higher level processes. By including these categories in your TOS you will be reminded to incorporate a wider range of cognitive processes into your tests.

The numbers in the body of the table reflect the number of items to be written to assess each content area at each level of the cognitive taxonomy. [Table 2.3](#) depicts specifications for a 35-item test. If you examine the first content area in [Table 2.3](#) (i.e., Measurement Error) you see that three knowledge level items, three comprehension level items, one analysis level item, and one synthesis level item are dedicated to assessing this content area. The next content area (i.e., Classical Test Theory) will be assessed by two knowledge level items, two analysis level items, and two synthesis level items. The number of items devoted to assessing each objective should reflect the importance of the objective in the curriculum and how much instructional time was devoted to it. Some test developers recommend using percentages instead of the number of items since the final item count might change during test development.

While TOSs were originally designed for achievement tests, they can be helpful with other maximum-performance tests and even typical-response tests. For example, anxiety is a multifaceted problem and if you want to develop a measure of anxiety you might want to ensure that you include items covering cognitive symptoms (e.g., worry), physiological symptoms (e.g., muscle tension), and behavioral symptoms (e.g., fidgeting). By including these as content areas in your TOS you can help ensure that your items cover the broad domain of symptoms of anxiety.

## ***Specify item format***

The next step is to determine the type of items you will include in your test. There is an abundance item formats available and it is beyond the scope of this chapter to review even the most common item formats, compare their strengths and weaknesses, and provide guidelines for item development. There are many resources that provide this information (e.g., [Reynolds & Livingston, 2012](#); [Reynolds, Livingston, & Willson, 2009](#)). However, we can outline some general principles that should be considered when determining what type of items to use on your test.

### ***Select the item format that most directly measures the construct***

First, the most important principle when selecting an item format is to select the format that provides the purest, most direct measurement of the construct. If you want to measure an examinee's writing ability, an essay item would likely be the most direct measure. If you want to measure an examinee's ability to pilot an airplane, a performance assessment utilizing an actual aircraft or a simulator might be

the format of choice. When assessing feelings, self-talk, and other covert behaviors, we often rely on self-reports using some type of selected-response format. But even here we have different formats to choose from, some with rather subtle differences. For example, many self-report scales use a simple True–False response option. However, self-report scales can also use rating scales. Take the item, “I feel anxious.” By changing the response format we change the interpretation of the responses. Consider these variations:

- I feel anxious. True/False.
- I feel anxious. Never/Sometimes/Often/Almost always.
- I feel anxious. Daily/Weekly/Monthly/Once a Year or Less/Never.

The first item asks more about a current state, that is, how you feel right now, while the other options look at longer-term trends in feelings. So even on a self-report scale, subtle changes in item formats influence how we interpret the responses. When selecting item format it is important to go back and ensure that you are being consistent with your original conceptual and operational definitions of the constructs.

### ***Select item formats that promote reliability and validity***

The second guiding principle is to select items formats that are likely to result in reliable scores and valid interpretations. For example, if the construct can be measured equally well with a selected-response item (e.g., multiple-choice item) and a constructed-response item (e.g., essay) we recommend using selected-response items since they can typically be scored in a more reliable manner. Additionally, since examinees can typically answer more selected-response items compared to constructed-response items in the same amount of time, selected-response items often allow you to sample the content domain more thoroughly and enhance the reliability of test scores. Both of these examples support the use of selected-response items, but in practice it is important to remember the previous principle. You need to consider initially what format will provide the most pure measure of the construct then factor in other considerations. For example, if you decide that constructed-response items such as essays are the most direct measure of the construct, you should then follow sound item development procedures to ensure psychometrically sound results (e.g., develop detailed scoring rubrics to enhance reliable scoring).

### ***Estimate how many items to include***

Next you need to estimate the number of items you think will be required to reliably measure the construct. The optimal number of items to include in your test is determined by a number of factors. These include:

1. *Available time.* The time required to administer the test varies considerably from test to test. If you are developing a comprehensive intelligence test it might be reasonable to allot one and a half hours for the average administration time. The results of intelligence tests

may be used to make important decisions that impact examinees, so this amount of time is warranted to ensure the results demonstrate excellent reliability and validity. In contrast, for a brief instrument designed to screen a large number of examinees for anxiety you would likely limit the number of items so the test can be completed in 10 or 15 min. In this situation, where more thorough assessment procedures are available to address concerns about individual examinees, you would not expect as strong psychometric properties as you would of a comprehensive intelligence test.

2. *Examinee characteristics.* The age and stamina of the specified examinees can also impact the number of items you include in your test. For example, with young and elderly clients you may need to limit your testing time to 30 min to maximize effort, motivation, and concentration. With adolescents and adults the testing period may potentially be expanded to two hours or more. In addition to age other examinee characteristics may influence the optimal length of your test. For example, examinees with certain disorders such as Attention Deficit Hyperactivity Disorder (ADHD) typically have limited attention spans, and unless the test is intended to measure attention, lengthy scales may be confounded by the attentional problems.
3. *Item format.* As noted previously, in most cases examinees can complete more selected-response items in a given time period than they can complete constructed-response items.

Clearly determining how many items to include in your test is a complex process involving a number of factors. While this is not an exhaustive list, it does highlight some salient factors to consider when determining how many items to include on your test. It should be noted that at this point you are still estimating how many items you will likely include in your test. When empirical analyses are available you might actually decide to include more or fewer items than initially anticipated.

## ***Plan for item development***

Once you have selected the item format(s) and estimated how many items to include in your test you should describe your plan for developing the items. First, specify who will actually develop the items. Unless your test measures a fairly narrow construct where you are a leading researcher, you will probably need assistance from content-area experts to write a comprehensive set of effective items. No one has the content expertise to write items that cover all aspects of achievement, all aspects of psychopathology, all aspects of temperament, etc. In addition to content knowledge you may also need professionals with extensive knowledge of the target population (e.g., developmental stages, ethnicity). When the time comes to actually write the items, the accepted rule of thumb is to initially write twice as many items as you anticipate using on the test. For constructs that are new or difficult to measure, three times the anticipated number of items is reasonable for the initial draft. You need to write a large number of items because many will drop out due to poor item statistics, redundancy, and concerns about bias.

Once you have specified the item formats you will include in your test, you should write sample items for each format. In writing items it is important to follow the established principles for developing psychometrically sound items. These principles are available in most comprehensive psychometric textbooks (e.g., [Reynolds](#)

& Livingston, 2012; Reynolds et al., 2009). When writing the sample items, also indicate the correctly keyed response (for select-response items) or provide sample responses (for constructed-response items).

This is also the time to write the instructions for administering and scoring of the test. We recommend you start by drafting the instructions for the examiner to follow when administering the test and then write the instructions to be provided to the examinee. Our experience is that initial drafts of instructions require considerable work. They are usually too long and overly detailed or too brief and missing important information. Keep them succinct and to the point and strive for clarity. Once your draft instructions are completed, we recommend you attempt to give and complete sample items strictly following the instructions. Continue refining and revising your instructions until an individual unfamiliar with the format can administer the sample items accurately by following the directions. The same process applies to writing instructions for scoring the test. To assess the accuracy of scoring, have several individuals score sample items and assess interrater reliability.

### ***Specify methods for item tryout and selection***

In this stage you should specify what procedures you will use to select the final set of items to include on your test. This is an extremely important stage and your decisions should be data-driven. The following steps are recommended:

#### ***Diversity panel review***

A panel of content-area experts representing different ethnic, gender, and religious groups should review the items. Research has shown that the use of diverse panels is not actually effective in detecting culturally biased items, but they can be helpful at this stage in identifying items that are offensive or ambiguous in specific cultures.

#### ***Specify a plan for item tryout***

Test developers typically collect a small sample of the target population and administer the draft test to them. This allows you to receive preliminary feedback from examiners and examinees on all aspects of the test, such as the clarity of instructions, ambiguity in item content and/or format, and scoring. This also allows you to statistically examine the items to detect any that are performing poorly at this early stage and either revise or eliminate them before moving forward with further data collection.

#### ***Specify the statistical methods you will use to select items***

There are a number of statistical procedures that allow you to empirically evaluate the properties of individual items. It is important to specify which procedures you will use to select items at this stage in the developmental process, as this will impact how you proceed with data collection. For example, some procedures (e.g.,

those based on Item Response Theory) require larger samples to produce stable results than other techniques. Similarly, some procedures are particularly sensitive to disproportional sample sizes across groups (e.g., male/female, ethnic groups) while others are more accommodating of these variations. By identifying what procedures you will use for selecting items for your test at this stage you can plan appropriately for subsequent data collection. Below is a brief description of the most common procedures.

1. *Item difficulty index.* The item difficulty index ( $p$ ) is defined as the percentage of examinees who correctly answer the item. For maximizing variance and reliability, a  $p$  of 0.50 is optimal (half of the examinees answer the item correctly). However, for a number of technical reasons (e.g., item intercorrelations) the general guideline is for your items to have a range of 0.20 around the optimal value (e.g., [Aiken, 2000](#)). For example, you might select items with  $p$  values ranging from 0.40 to 0.60 with a mean of 0.50. While a mean  $p$  of 0.50 is considered optimal in some applications, different testing applications require different levels. For example, to take into consideration the effects of random guessing, the optimal item difficulty level is set higher for selected-response items. [Lord \(1952\)](#) recommends that for multiple-choice items with four options the average  $p$  value should be approximately 0.74. In this context you would select items with  $p$  values ranging from 0.64 to 0.84 with a mean of approximately 0.74. The item difficulty index is only applicable with maximum-performance tests where items are scored correct/incorrect, but on typical-response tests a similar index, the percent endorsement, may be calculated (you may also see this referred to as the Attractiveness Index). The percent endorsement index indicates the percentage of examinees responding in a given manner (e.g., indicating the presence of a trait). It is important to note that both the item difficulty index and the percent endorsement index are sample-dependent statistics. A math item that is difficult in a sample of students in the third grade would likely be easier in a sample of students in the sixth grade. Similarly, an item assessing depression might be frequently endorsed in a clinical sample but rarely endorsed in a community sample.
2. *Item discrimination.* Item discrimination refers to how well an item discriminates between examinees that differ on the construct being measured. Obviously discrimination is of paramount importance when selecting items for your tests. There have been over 50 indices of discriminating power developed; however, they provide similar information ([Anastasi & Urbina, 1997](#)). We will briefly describe two of these indices. First, the *item-discrimination index* ( $D$ ) denotes the difference between the percentages of high and low scorers who gave the keyed response. The formula shown below yields this index.

$$D_i = \text{item-discrimination index}$$

$$D_i = \frac{n_{hi}}{n_h} - \frac{n_{li}}{n_l}$$

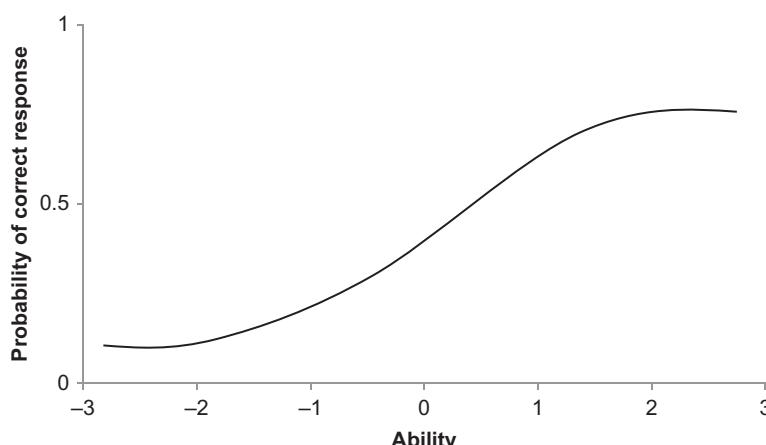
$n_{hi}$ , number of examinees in the high-scoring group who answered  $i$  correctly;  $n_h$ , number of examinees in the high-scoring group;  $n_{li}$ , number of examinees in the low-scoring group who answered item  $i$  correctly;  $n_l$ , number of examinees in the low-scoring group.

This notation uses correct/incorrect scoring, but the formula can be easily applied to typical-response tests. The high and low-groups can be defined in different ways, but they are typically defined in terms of total test performance. One common approach is to select the top and bottom 27% of examinees in terms of overall test performance and exclude

the middle 46% (Kelly, 1939). Hopkins (1988) provided guidelines for interpreting  $D$  values, describing  $D$  values of 0.40 as excellent, between 0.30 and 0.39 as good, between 0.11 and 0.29 as fair, and between 0.00 and 0.10 as poor. Items with negative  $D$  values are miskeyed or have other serious problems. We have slightly more rigorous standards and believe that items with  $D$  values less than 0.30 should be carefully reviewed and possibly revised or deleted.

The second approach to examining item discrimination which we will discuss is the use of item–total correlations. When items are scored in a dichotomous manner (e.g., correct/incorrect), the item–total correlation is usually calculated using the point–biserial correlation. The total score can be either total number of items answered in the keyed direction (unadjusted) or total number of items omitting the item being examined (adjusted). A strong item–total correlation is interpreted as evidence that the item under examination is measuring the same construct as the overall test and discriminates well between examinees high on the construct and those low on the construct. The results of an item–total correlation are comparable to that of the discrimination index and can be interpreted in a similar manner (Hopkins, 1988). Like the item difficulty index, the item-discrimination index and the item–total correlation are sample-dependent statistics and may vary from sample to sample.

3. *Item Characteristic Curves and Item Response Theory.* The item analyses we have described up to this point have been around for close to a century and were developed in the context of Classical Test Theory (CTT). While these techniques are well established and useful, they are often complemented by new techniques associated with Item Response Theory (IRT). Central to IRT is a mathematical model that describes how examinees at different levels of ability (or whatever latent trait is being assessed) respond to individual test items. While a discussion of IRT is well beyond the scope of this chapter, one component of IRT, the Item Characteristic Curve (ICC), is germane. ICC is a graph with ability reflected on the horizontal axis and the probability of a correct response reflected on the vertical axis. Each item has its own specific ICC, and these ICCs are plotted from mathematically derived functions and usually involve iterative procedures (Anastasi & Urbina, 1997). Fig. 2.1 presents a hypothetical—but typical—ICC for an



**Figure 2.1** Common item characteristic curve.

item of moderate difficulty and with good discrimination. This prototypical ICC takes a “Lazy S” shape and it is an asymptote. This ICC indicates that low levels of ability are associated with a low probability of a correct response and as the ability level increases the probability of a correct response increases. In other words, examinees with greater ability have an increased probability of answering the item correctly, relative to those with lower ability. This is exactly what you would expect with an item that is performing well.

ICCs incorporate information about the measurement characteristics of items that we have already discussed: the item's difficulty and discrimination ability. In ICC jargon the midpoint between the lower and upper asymptotes is referred to as the inflection point and reflects the difficulty of the item. The inflection point identifies the ability level at which an examinee has a 50% chance of getting the item correct. Discrimination is demonstrated by the slope of the ICC at the inflection point. ICCs with steeper slopes demonstrate better discrimination than those with gentler slopes. Fig. 2.2 illustrates how both difficulty and discrimination are reflected in an ICC. The vertical dotted line marked “Difficulty” indicates that an examinee needs a z-score of approximately 0.33 to have a 50% chance of answering the item correctly. The dotted line marked “Discrimination” is at about a 45 degree angle, which suggests good discrimination.

IRT provides some advantages over the item analyses developed in CTT. As noted, item difficulty and item discrimination in CTT are sample dependent. That is, the statistics may vary from sample to sample. In IRT item difficulty and item discrimination are sample independent. That is, they are invariant across samples. This property of IRT parameters makes them particularly helpful in some applications such as computer adaptive testing (CAT) and developing alternate tests forms.

Another important application of IRT is the detection of biased items. In this context ICCs for different groups can be calculated (e.g., male/female). These ICCs can then be statistically analyzed to determine the degree of differential item

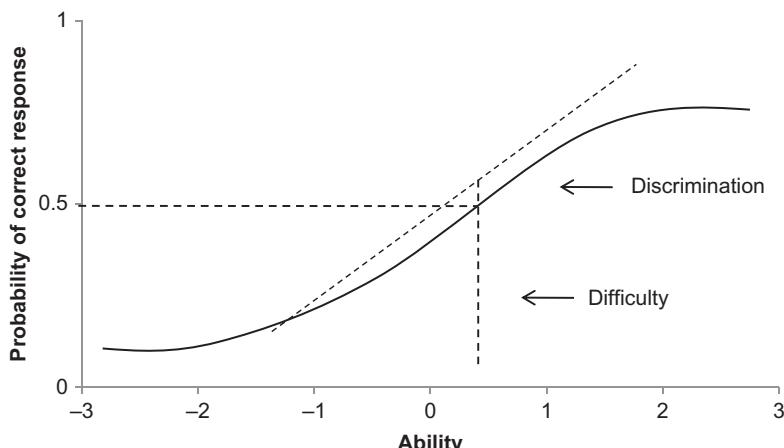
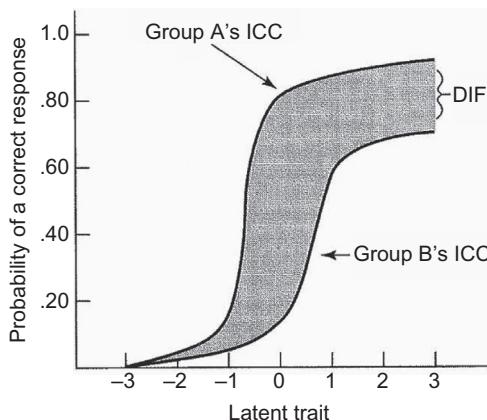


Figure 2.2 Item characteristic curve demonstrating difficulty and discrimination.



**Figure 2.3** Two ICCs illustrating differential item function (DIF).

**Table 2.4** Phase III—standardization, scaling, and psychometric studies

- |                                      |
|--------------------------------------|
| Step 1: Specify/standardization plan |
| Step 2: Specify scaling methods      |
| Step 3: Describe reliability studies |
| Step 4: Describe validity studies    |

function (DIF). [Embreton and Reise \(2000\)](#) provide a good introduction to IRT, including an approachable chapter on detecting DIF. The use of IRT combined with a logical analysis of item content is currently the most common method for detecting biased items ([Fig. 2.3](#)).

In this section we described some of the most common approaches to examining the measurement properties of individual test items. These will provide important information to help you select items to include in your test. However, this discussion was not exhaustive and other analyses may prove useful.

## Phase III: Planning standardization, scaling, and psychometric studies

In this phase you should describe both the standardization procedures that will be used and the scaling methods, specify the reliability and validity studies planned, and describe any special studies that might be required. [Table 2.4](#) provides an outline of these steps.

### ***Specify standardization plan***

In Phase II we encouraged you to specify the age range of the examinees your test will target, which was actually an initial step in describing the target population of

your test, which in turn sets the parameters for your standardization sample or reference group. For tests producing norm-referenced score interpretations the standardization sample forms the basis for normative scores (also known as norms) and serves as the reference group to whom examinees are compared. We actually have a strong preference for using the term reference group for test score interpretation, but continue to use “standardization sample” as the term is so common in testing and assessment works. Test users and developers should keep in mind that all such samples are no more than groups against which we reference or refer the performance or scores of a test-taker in order to clarify the meaning of the score (more will be said on this issue in the scaling section below). In tests producing criterion-referenced score interpretations, performance of reference samples will be important in designating cut scores and related interpretive decisions and the designation of such samples as true reference groups seems clearer. Once you have defined the target population you next develop a sampling plan.

The goal of your sampling plan is to ensure that your sample is representative of a relevant target population or reference group designed to answer questions posed about the examinee. The most common practice is to specify a population proportionate stratified random sampling plan when examinees are to have their performance compared to that of the population of a specified country. To accomplish this, you first determine what characteristics (strata) of the population are important for ensuring the sample is representative. In addition to age range you might consider other variables such as primary language, gender, ethnic background, education, socioeconomic status, geographic region, and population density (e.g., urban vs rural). The best source of information on population demographics is the United States Bureau of the Census. The overall goal is to sample individuals from all demographic groups according to their relative proportions in the population at large. For example, if 1% of our target population are African American men between the ages of 30 and 35 years living in the northeast region of the United States in cities with a population of more than 250,000 and with an educational level of a bachelor’s degree, you would want 1% of the standardization sample to include individuals who match these characteristics.

While a nationally representative standardization sample is the most common reference group, other samples may be appropriate for specific applications. For example, school districts often provide local normative data for achievement tests that are based on the performance of students in their school district. This ensures that their students are compared with students who are similar on important variables. Some tests, in addition to nationally representative norms, provide “clinical” norms. For example, the Behavior Assessment System for Children, Second Edition (BASC-2: [Reynolds & Kamphaus, 2004](#)) includes normative data based on clinical samples in addition to general population norms. These clinical norms may be useful in refining diagnostic impressions. If special normative data are reported it is important to describe carefully these reference groups so users of the test can make informed decisions regarding how appropriate the data are for their application. In essence, different normative samples provide different information and address different questions.

You also need to specify the appropriate size of the standardization sample. The size of the sample is impacted by a number of factors. The most salient factor to consider is how the test results will be used. When test results can significantly impact individual examinees (e.g., personnel selection, clinical diagnosis) larger samples are required compared to tests that are used for preliminary screening or pure research applications. Since some statistical analyses require larger samples for stable results than do others, the statistical analyses you intend to use will also impact the size of your standardization sample. Finally, practical issues such as how much time and money is available for data collection will also come into play.

### ***Specify scaling methods***

Just as different standardization samples or reference groups provide different information and allow us to answer different questions, different types of scores also provide different information and allow us to address different questions. Most scores can be classified as raw scores, norm-referenced scores, criterion-referenced scores, or IRT-based scores. However, before dealing with these types of scores, it is necessary to review scaling and scales of measurement.

### ***Scales of measurement***

Measurement is a set of rules for assigning numbers to represent objects, traits, attributes, or behaviors. Psychological tests are measuring devices, and as such they involve rules (e.g., specific items, administration, and scoring instructions) for assigning numbers to an individual's performance that are interpreted as reflecting characteristics of the individual. When we measure something, the units of measurement have a mathematical property called the scale of measurement. A scale is a system or scheme for assigning values or scores to the characteristic being measured. Stevens (1946) originally proposed a taxonomy that specified four scales of measurement which continues to be useful to this day. These different scales have distinct properties and convey unique types of information. The four scales of measurement are nominal, ordinal, interval, and ratio. The scales form a hierarchy, and as we progress from nominal to ratio scales we are able to perform increasingly sophisticated measurements that capture more detailed information.

#### **Nominal scales**

Nominal scales are the simplest of the four scales. Nominal scales provide a qualitative system for categorizing people or objects into categories, classes, or sets. In most situations, these categories are mutually exclusive. For example, sex is an example of a nominal scale that assigns individuals to mutually exclusive categories. Another example is assigning people to categories based on their college academic major (e.g., psychology, biology, chemistry). Numbers can be assigned to nominal scales simply to identify or label the categories; however the categories are not ordered in a meaningful manner and are not subject to arithmetical manipulation in any meaningful way. The assignment of numbers in ordinal scaling is

completely arbitrary and as such nominal scales do not actually quantify the variables under examination. Numbers assigned to nominal scales should not be added, subtracted, ranked, or otherwise manipulated. As a result, most common statistical procedures cannot be used with these scales, so their usefulness is limited.

### Ordinal scales

Ordinal scale measurement allows you to rank people or objects according to the amount or quantity of a characteristic they display or possess. As a result, ordinal scales enable us to quantify the variables under examination and provide substantially more information than nominal scales. For example, ranking people according to height from the tallest to the shortest is an example of ordinal measurement. Traditionally, the ranking is ordered from the “most” to the “least.” In our example the tallest person in the class would receive the rank of 1, the next tallest a rank of 2, and so on. Although ordinal scale measurement provides quantitative information, it does not ensure that the intervals between the ranks are consistent. That is, the difference in height between the persons ranked 1 and 2 might be three inches while the difference between those ranked 3 and 4 might be one inch. Ordinal scales indicate the rank-order position among individuals or objects, but they do not indicate the extent to which they differ. All the ordinal scale tells us then is who is taller, number 5 or number 7; it tells us nothing about how much taller. As a result, these scales are significantly limited in both the measurement information they provide and the statistical procedures that can be applied. Although you will see it done, it rarely makes sense to add, subtract, multiply, or divide such scores or to find their mean. Nevertheless, the use of these scales is fairly common in many settings. Percentile ranks, age-equivalents, and grade-equivalents are all examples of common ordinal scales.

### Interval scales

Interval scales provide more information than either nominal or ordinal scales. Interval scale measurement allows you to rank people or objects like an ordinal scale, but on a scale with equal units. By equal scale units, we mean the difference between adjacent units on the scale is equivalent. The difference between scores of 70 and 71 is the same as the difference between scores of 50 and 51 (or 92 and 93; 37 and 38; etc.). Many psychological tests are designed to produce interval-level scores. Let's look at an example of scores for three people on an intelligence test. Assume individual A receives a score of 100, individual B a score of 110, and individual C a score of 120. First, we know that person C scored the highest followed by B then A. Second, given that the scores are on an interval scale, we also know that the difference between individuals A and B (i.e., 10 points) is equivalent to the difference between B and C (i.e., 10 points). Finally, we know the difference between individuals A and C (i.e., 20 points) is twice as large as the difference between individuals A and B (i.e., 10 points). Interval-level data can be manipulated using common mathematical operations (e.g., addition, subtraction, multiplication, and division) whereas lesser scales (i.e., nominal and ordinal) cannot. A final advantage is that most statistical procedures can be used with interval scale data.

Interval scales represent a substantial improvement over ordinal scales and provide considerable information. Their major limitation is that interval scales do not have a true zero point. That is, on interval scales a score of zero does not reflect the total absence of the attribute. For example, if an individual were unable to answer any questions correctly on an intelligence test and scored a zero, it would not indicate the complete lack of intelligence, but only that they were unable to respond correctly to any of the questions in this test. Additionally, ratios are not meaningful with interval scales. For example, even though an IQ of 100 is twice as large as an IQ of 50, it does not mean that the person with a score of 100 is twice as intelligent as the person with a score of 50. For such a statement to be accurate, we would need to have a true zero point.

Despite this limitation, some school districts and agencies continue to use some form of a “percentage discrepancy” between actual and predicted achievement or an IQ. In such a circumstance the difference between two values, such as an obtained achievement score of say 75, is subtracted from the student’s predicted achievement score of 100. The difference score is then used to calculate a percentage of deficiency in the area of academics covered. Various formulas are in use to do this, but, simplistically, one might take this difference of 25 points, divide it by 100 (the predicted score), and decide the student has a 25% deficiency in the area in question. Such a percentage is nonsensical, regardless of the formulas used to make the determination, because a percentage is a ratio of two numbers, and ratios are uninterpretable in interval scaling because we have no true zero point for reference in interval scaling. Ratio scales are required, as described below.

With behavioral variables like intelligence, or even personality characteristics like friendliness, we do not know where the true zero point lies. With physical characteristics like height and weight, a zero point is well defined, and we measure beginning at zero and go up. When the zero point is unknown, the only place we can begin measuring accurately is the middle of the distribution. Interval scales are derived by first locating the midpoint of a variable in some defined population or reference group, usually taken to be the mean score, and then measuring outward in each direction, above and below, as far as we can establish scores with reasonable accuracy. We never reach the true bottom or true top of what we are measuring (although a particular test may bottom-out or top-out, the construct continues). Remember that interval scales, the most common scale in psychology and education, begins measuring in the middle—which is the only point we can initially define—and then measures towards the two ends, or tails, of the distribution, never reaching either end (because the normal curve is asymptotic to its axis).

### Ratio scales

Ratio scales have the properties of interval scales plus a true zero point that reflects the complete absence of the characteristic being measured. Miles per hour, length, and weight are all examples of ratio scales. As the name suggests, with these scales we can interpret ratios between scores. For example, 100 miles/h is twice as fast as 50 miles/h, 6 feet is twice as long as 3 feet, and 300 pounds is three times as much as 100 pounds. Ratios are not meaningful or interpretable with other scales. As we

noted, a person with an intelligence quotient (IQ) of 100 is not twice as intelligent as one with an IQ of 50. Given the enormity of human intelligence, an IQ of 100 may only represent a 1%, 5%, or 10% increase over an IQ of 50. The key point being that absent a ratio scale for intelligence, we cannot know the magnitude of such a difference in absolute terms. This holds in achievement as well. A person with a standardized math achievement test score of 120 does not know “twice as much” as one with a score of 60. A person with a psychopathy score of 70 may or may not be “twice as psychopathic” as a person with a score of 35. With the exception of the percentage of items that are correct and the measurement of behavioral responses (e.g., reaction time), there are relatively few ratio scales in psychological measurement. Fortunately, we are able to address most of the measurement issues in psychology adequately using interval scales.

As we noted, there is a hierarchy among the scales with nominal scales being the least sophisticated and providing the least information and ratio scales being the most sophisticated and providing the most information. Nominal scales allow you to assign a number to a person which associates that person with a set or category, but other useful quantitative properties are missing. Ordinal scales have all the positive properties of nominal scales with the addition of the ability to rank people according to the amount of a characteristic that they possess. Interval scales have all the positive properties of ordinal scales and also incorporate equal scale units. The inclusion of equal scale units allows one to make relative statements regarding scores (e.g., the difference between a score of 82 and a score of 84 is the same as the difference between a score of 92 and 94). Finally, ratio scales have all of the positive properties of an interval scale, with the addition of an absolute zero point. The inclusion of an absolute zero point allows us to form meaningful ratios between scores (e.g., a score of 50 reflects twice the amount of the characteristic as a score of 25). Although these scales do form a hierarchy, this does not mean the lower scales are of little or no use. If you want to categorize students according to their academic major, a nominal scale is clearly appropriate. Accordingly, if you simply want to rank people according to height then an ordinal scale would be adequate and appropriate. However, in most measurement situations you want to use the scale that provides the most information. Once you have determined the appropriate scale of measurement for a test, various scores can be derived and created with a variety of reference groups, but do keep in mind that the type of score itself and the reference group involved also are keys to what questions a score can help you answer, and all must be appropriate to the task—not just one.

The following are brief descriptions of the major types of scores:

*Raw scores:* Raw scores are simply the number of items scored in a specific manner, such as correct/incorrect, true/false, etc. Raw scores typically reflect ordinal scale measurement and provide limited information when interpreting test results. Since raw scores often have limited interpretive value, they are commonly transformed into another score format to facilitate interpretation.

*Norm-referenced score interpretations:* With norm-referenced score interpretations the examinee’s performance is compared to the performance of other examinees (typically those in the standardization sample or normative group). The most

common norm-referenced scores are standard scores, which are sometimes referred to as scaled scores. Transforming raw scores into standard scores involves creating a distribution with a predetermined mean and standard deviation that remains constant across some preselected variable such as age. This is termed “scaling” because we change the underlying metric or rescale the scores. All standard scores use standard deviation units to indicate where an examinee’s score is located relative to the mean of the distribution. Standard score formats differ in their means and standard deviations. The most common standard score formats are *z*-scores (mean of 0 and a standard deviation of 1), *T*-scores (mean of 50 and a standard deviation of 10) and IQs (mean of 100 and a standard deviation of 15). Standard scores are generally considered to reflect interval-level measurement.

When variables that are normally distributed (or approximate normality) standard scores are typically computed using a linear transformation. Standard scores calculated using linear transformation retain a direct relationship with the original raw scores and the shape of the distribution is unchanged. When variables deviate markedly from the normal distribution test developers often develop normalized standard scores. Here the original distribution, which was not normal, is transformed into a normal distribution, often using nonlinear transformations that fit scores to the approximate shape of the so-called normal or Gaussian curve. When normalized standard scores are developed using nonlinear transformations the scores may not retain a direct relationship with the raw scores and the shape of the true distributions may change. The use of nonlinear transformations is not necessarily undesirable depending on why the distribution was not normal to start with. If the variable is not normally distributed in the natural world, normalization is usually not warranted and often may be misleading. However if the variable is normally distributed and minor deviations from normality are the result of sampling error, normalization may be useful and enhance interpretation. If the rationale for normalization is sound, normalized standard scores can be interpreted in the same manner as other standard scores. Wechsler scales scores (mean of 10 and standard deviation of 3) and normal curve equivalents (NCE; mean on 50 and standard deviation of 21.06) have traditionally been based on nonlinear transformation and are normalized standard scores.

The percentile rank is another norm-referenced score and is one of the most easily understood ways to report and interpret test results. Like all norm-referenced scores, the percentile rank simply reflects an examinee’s performance relative to a specific group. Percentile ranks are interpreted as reflecting the percentage of individuals scoring below a given point in a distribution. For example, a percentile rank of 70 indicates that 70% of the individuals in the standardization sample scored below this score. A percentile rank of 20 indicates that only 20% of the individuals in the standardization sample scored below this score. Percentile ranks range from 1 to 99, and a rank of 50 indicates the median performance. Percentile ranks can be explained easily to and understood by individuals without formal training in psychometrics, and are often useful when explaining test results to examinees. Although percentile ranks can be interpreted easily, they represent ordinal scale measurement. That is, percentile ranks do not have equal intervals across a

distribution. Percentile ranks are compressed near the middle of the distribution, where there are large numbers of scores, and they are spread out near the tails where there are relatively few scores. As a result small differences in percentile ranks near the middle of the distribution might be minimal, whereas the same differences at the extreme ends of the distribution might be substantial. Therefore it is often noted that use of percentile ranks will exaggerate small differences near the mean and obscure large differences near the tails of the distribution. However, since the pattern of inequality is predictable and well known, this can be taken into consideration when interpreting scores and it is not problematic as long as the user is aware of this characteristic of percentile ranks.

Grade equivalents are norm-referenced scores that identify the academic “grade level” achieved by the examinee. While grade equivalents are popular in some settings and appear to be easy to interpret, they need to be interpreted with considerable caution. Much has been written about the limitations of grade equivalents, and while we would not go into detail about these limitations in this chapter, we do want to emphasize that they are subject to misinterpretation and should be avoided when possible (for a full discussion see [Reynolds & Livingston, 2012](#)). Age equivalents share many of the limitations of grade equivalents and should also be avoided when possible. Grade and age equivalents are ordinal level scales of measurement, and thus are not subject to the manipulations allowed with interval-level scaling.

*Criterion-referenced score interpretations:* With criterion-referenced score interpretations, the examinee’s performance is not compared to the performance of other examinees; instead it is compared to a specified level of performance (i.e., a criterion). With criterion-referenced interpretations, the emphasis is on what the examinees know or what they can do, not their standing relative to other examinees. One of the most common examples of a criterion-referenced score is the percentage of correct responses on a classroom examination. If you report that a student correctly answered 75% of the items on a classroom test, this is a criterion-referenced interpretation. Notice that you are not comparing the examinee’s performance to that of other examinees; instead you are comparing their performance to a standard or criterion, in this case perfect performance on the test.

Another popular criterion-referenced interpretation is referred to as *mastery testing*. Mastery testing involves determining whether the examinee has achieved a specific level of mastery of the knowledge or skills domain and is usually reported in an all-or-none score such as a pass/fail designation ([AERA et al., 2014](#)). The written exam required to obtain a driver’s license is an example of a mastery test. The test is designed to determine whether the examinee has acquired the basic knowledge necessary to operate a motor vehicle successfully and safely (e.g., state motorizing laws and standards). A *cut score* had been previously specified and all scores equal to or above this score are reported as “pass” whereas scores below it are reported as “fail.”

A final criterion-referenced interpretative approach is referred to as “standards-based interpretations.” Whereas mastery testing typically results in an all-or-none interpretation (i.e., the examinee either passes or fails), standards-based interpretations usually involve three to five performance categories. For example, the results

of an achievement test might be reported as basic, proficient, or advanced. Another variation of this approach is the assignment of letter grades to reflect performance on classroom achievement tests. For example, A's might be assigned for percentage correct scores between 90% and 100%, B's for scores between 80% and 89%, C's for scores between 70% and 79%, D's for scores between 60% and 69%, and F's for scores below 60%.

*Scores based on IRT:* Theoretical and technical advances have ushered in new types of scores that are based on IRT. In brief, IRT is a theory of mental measurement that holds that the responses to test items are accounted for by latent traits (see [Reynolds & Livingston, 2012](#) for a gentle introduction to IRT). In IRT it is assumed that each examinee possesses a certain amount of any given latent trait, and the goal is to estimate the examinee's ability level on the latent trait. The specific ability level of an examinee is defined as the level on a scale where the examinee can correctly respond to half of the items. In IRT terminology, an examinee's ability level is designated by the Greek letter theta ( $\theta$ ). The scores assigned to reflect an individual's ability level in IRT models are similar to the raw scores on tests developed using traditional models (i.e., Classical Test Theory). That is, they can be transformed into either norm or criterion-referenced scores. However, IRT scores have a distinct advantage over traditional raw scores; they are interval-level scores they and have stable standard deviations across age groups. These IRT scores have different names, including *W*-scores, growth scores, and Change Sensitive Scores (CSS). Some refer to them generically as Rasch or Rasch-type scores after the originator of the mathematical models. *W*-scores are used on the Woodcock–Johnson III ([Woodcock, McGrew, & Mather, 2001](#)) and are set so a score of 500 reflects cognitive performance at the beginning fifth grade ability level. *W*-scores have proven to be particularly useful in measuring changes in cognitive abilities. For example, they can help measure gains in achievement due to learning or declines in cognitive abilities due to dementia. In terms of measuring gains, if over time an examinee's *W*-score increases by 10 units (e.g., from 500 to 510), they can now complete tasks with a 75% probability of success that they originally could only complete with a 50% probability of success. Conversely, if an examinee's *W*-score decreases by 10 units (e.g., from 500 to 490), they can now complete tasks with only a 25% probability of success that they originally could complete with a 50% probability of success ([Woodcock, 1999](#)).

What scores should be used? As noted, different types of scores provide different information and answer different questions. In order to answer the question of which scores to use, remember what information the test scores provide.

- Raw scores tell us the number of points accumulated by an examinee on a test. If we know the raw scores of all examinees they can also tell us the relative rank among examinees. Raw scores typically provide only ordinal scale measurement.
- Norm-referenced standard scores address the general question of how this examinee's performance compares to the performance of some specified reference group. Standard scores typically reflect interval scale measurement.
- Criterion-referenced scores tell us whether or not or an examinee's performance has reached a desired level of proficiency.

- Rasch or IRT-based scores are on an equal interval scale and reflect standing on an underlying or latent trait. These scores are particularly useful in evaluating the degree of change in scores over time and in comparing scores across tests of a common latent trait.

This illustrates that each type of score provides us with a different type of information. Which score we should use is dependent upon the type of information your test provides. It should be noted that you do not need to limit your test to producing only one type of score. It is not uncommon for tests developed using CTT to report both norm- and criterion-referenced scores. If you use IRT in developing your test you can report IRT-based scores along with norm- and criterion-referenced scores.

### ***Specify reliability/precision studies***

At this point you should specify the reliability studies you will perform. In the context of measurement, reliability refers to the stability, accuracy, or consistency of scores produced by a measurement. Errors of measurement undermine the reliability of measurement and therefore reduce the usefulness of the test results. There are multiple sources of measurement error with the two major sources being content-sampling errors and time-sampling errors. Content-sampling errors are the result of less than perfect sampling of the content domain. The error that results from differences between the sample of items (i.e., the test) and the domain of items (i.e., all the possible items) is referred to as content-sampling error. If the items on a test are a good sample of the domain, the amount of measurement error due to content sampling will be relatively small. If the items on a test are a poor sample of the domain, the amount of measurement error due to content sampling will be relatively large. Content-sampling error is typically considered the largest source of error in test scores and therefore is the source that concerns us most. Measurement error can also be introduced by one's choice of a particular time to administer the test. This type of measurement error is referred to as a time-sampling error and reflects random fluctuations in performance from one situation or time to another, and also limits our ability to generalize test results across different situations. Although errors due to content sampling and time sampling typically account for the largest proportion of random error in testing, administrative and scoring errors that do not affect all examinees equally will also contribute to the random error observed in scores. For example, when the scoring of a test relies heavily on the subjective judgment of the person grading the test it is important to consider differences in graders, usually referred to as interrater or interscorer differences. That is, would the examinee receive the same score if different individuals graded the test?

In CTT, reliability coefficients can be classified into three broad categories (AERA et al., 2014). These include: (1) coefficients derived from the administration of the same test on different occasions (i.e., test-retest coefficients), (2) coefficients based on the administration of parallel forms of a test (i.e., alternate-form coefficients), and (3) coefficients derived from a single administration of a test (i.e., internal-consistency coefficients). A fourth type, interrater reliability, is indicated when scoring involves a significant degree of subjective judgment. We will briefly describe these methods of estimating reliability.

### ***Test–retest coefficients***

Possibly the most straightforward way to estimate the reliability of a test score is to administer the same test to the same group of examinees on two different occasions and then calculate the correlation between the scores on the two administrations. The result is a test–retest coefficient and is primarily sensitive to time-sampling error. It is an index of the stability of test scores over time, and some authors refer to coefficients obtained with this approach as stability coefficients. Test–retest analyses also allow you to evaluate practice or carry-over effects. This information is valuable when interpreting scores for the same examinee taken on different occasions.

### ***Alternate-form coefficients***

If two equivalent or parallel forms of the test are available both forms of the test can be administered to the same group of examinees and a correlation calculated between the scores on the two assessments. Two fairly common procedures are used to establish alternate-form reliability. One is alternate-form reliability based on simultaneous administrations and is obtained when the two forms of the test are administered on the same occasion (i.e., back-to-back). The other, alternate form with delayed administration is obtained when the two forms of the test are administered on two different occasions. Alternate-form reliability based on simultaneous administration is primarily sensitive to content-sampling error. Alternate-form reliability with delayed administration is sensitive to measurement error due to both content-sampling and time-sampling error.

### ***Internal consistency coefficients***

Internal-consistency coefficients are based on the relationship between items within a test, are derived from a single administration of the test, and primarily reflect content-sampling error. Split-half reliability coefficients are calculated by administering a test to a group of examinees and then dividing the test into two equivalent halves that are scored independently. The results on one half of the test are then correlated with results on the other half of the test. While there are many ways a test can be divided in half the most common approach is to use an odd–even split. Since this approach is actually correlating two halves of the test, researchers use the Spearman–Brown correction formula to provide an estimate of the reliability of the full or whole test.

Other internal-consistency coefficients are based on formulas developed by Kuder and Richardson (1937) and Cronbach (1951). Instead of comparing responses on two halves of the test as in split-half reliability, this approach examines the consistency of responding to all the individual items on the test. These reliability estimates can be thought of as the average of all possible split-half coefficients and are properly corrected for the length of the whole test. Like split-half reliability, these estimates are primarily sensitive to content-sampling error. Additionally, they are sensitive to the heterogeneity of the test content. When we refer to content

heterogeneity we are referring to the degree to which the test items measure related characteristics. In their original article [Kuder and Richardson \(1937\)](#) presented numerous formulas for estimating reliability, the most common formula is known as the Kuder–Richardson Formula 20 (KR 20), which is applicable when test items are scored dichotomously. Coefficient alpha ([Cronbach, 1951](#)) is a more general form of KR 20 that also deals with test items that produce scores with multiple values (e.g., 0, 1, or 2). Because Coefficient alpha is more broadly applicable, it has become the preferred statistic for estimating internal consistency.

### ***Interrater reliability***

If the scoring of a test involves subjective judgment it is important to evaluate the degree of agreement between different individuals scoring the test. This is referred to as interscorer or interrater reliability. To estimate interrater reliability the test is administered one time and two individuals independently score each test. A correlation is then calculated between the scores obtained by the two scorers. This estimate of reliability primarily reflects differences due to the individuals scoring the test and is less sensitive to error due to content or time sampling. In addition to the correlational approach, interrater agreement can also be evaluated by calculating the percentage of times that two individuals assign the same scores to the performances of examinees. This approach is typically referred to as interrater agreement or percent agreement. Many authors prefer Cohen's kappa over the standard percent of agreement when analyzing categorical data as kappa is a more robust measure of agreement as it takes into consideration the degree of agreement expected by chance ([Hays, 1994](#)).

When describing CTT reliability coefficients we specified what source of error they were primarily sensitive to, but in actual practice more than one source of measurement error is actually reflected in any coefficient. For example, we noted that internal consistency coefficients are primarily sensitive to content-sampling error, which is accurate. However they are also influenced by administrative errors, scoring errors, and time-sampling errors to some degree. To overcome this, generalizability theory ([Cronbach, Rajaratnam, & Gleser, 1963](#)) allows the researcher to estimate the specific variance resulting from content-sampling error, time-sampling error, interrater differences, and the interaction of these factors. Generalizability theory typically uses analysis of variance (ANOVA) to estimate the variance component associated with each source and generalizability coefficients can be then calculated. As a result, while most test manuals report information on reliability based on CTT, generalizability theory does provide a more informative approach to reliability analyses.

In IRT information on the reliability of scores is typically reported as a test information function (TIF). A TIF reflects the reliability of measurement at different points along the distribution. A common finding is that the reliability of measurement is not constant across the distribution, often with the most reliable information around the middle of the distribution and less reliable information near the tails of the distribution. The ability to provide information about reliability at

different points of the distribution is an advantage of the TIF; however it does not provide information about generalizability across time or situations (AERA et al., 2014).

As a general rule we recommend that you perform reliability analyses that allow you to address both content-sampling and time-sampling error. For example, you might elect to report test-retest coefficients and Cronbach's alpha. If you develop alternate forms of your test you would naturally want to report alternate-form coefficients (ideally based on both simultaneous and delayed administrations). Finally, if there is any subjective judgment involved in scoring then interrater reliability analyses should be performed.

### **Specify validity studies**

It is also important to specify the validity studies in advance. *The Standards for Educational and Psychological Testing* (AERA et al., 2014) define validity as "...the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of the tests" (p. 11). The *Standards* specify five basic categories of validity evidence. These are:

#### ***Evidence based on test content***

Validity evidence can be garnered by examining the relationship between the content of the test and the construct or domain the test is designed to measure (AERA et al., 2014). Collecting content-based validity evidence is often based on the evaluation of expert judges about the correspondence between the test's content and the construct it is designed to measure. The key issues addressed by these expert judges are whether the test items assess relevant content (i.e., item relevance) and the degree to which the construct is assessed in a comprehensive manner (i.e., content coverage).

#### ***Evidence based on relations to other variables***

Validity evidence can also be collected by examining the relationship of test scores to other variables (AERA et al., 2014). In describing this type of validity evidence, the Standards recognize several related but distinct applications. These are:

- *Test-criterion evidence.* Many tests are designed to predict performance on some variable that is typically referred to as a criterion. The criterion can be academic performance as reflected by the grade point average (GPA), job performance as measured by a supervisor's ratings, or anything else that is of importance to the test user. Test-criterion validity evidence is often reported in terms of a validity coefficient.
- *Convergent and discriminant evidence.* Convergent evidence of validity is obtained when you correlate your test with existing tests that measure the same or similar constructs. Discriminant evidence of validity is obtained when you correlate your test with existing tests that measure dissimilar constructs. If your analyses produce the expected correlations, this can be reported as validity evidence for your proposed interpretations.

- *Contrasted groups studies.* Validity evidence can also be garnered by examining different groups, which are expected, based on theory, to differ on the construct the test is designed to measure. For example, do tests scores distinguish between individuals with an anxiety disorder and those with depression. If you find the expected relationships this can support the validity of your proposed score interpretation.
- *Evidence based on internal structure.* By examining the internal structure of a test (or battery of tests) one can determine whether the relationships among test items (or, in the case of test batteries, component tests) are consistent with the construct the test is designed to measure (AERA et al., 2014). Factor analysis is often used to determine the number of conceptually distinct factors or dimensions underlying a test or battery of tests. Factor analysis is not the only approach researchers use to examine the internal structure of a test and any technique that allows researchers to examine the relationships between test components can be used in this context.
- *Evidence based on response processes.* Validity evidence based on the response processes invoked by a test involves an analysis of the fit between the performance and actions the examinees engage in and the construct being assessed. There are numerous ways of collecting this type of validity evidence, including interviewing examinees about their response processes and strategies, recording behavioral indicators such as response times and eye movements, or even analyzing the types of errors committed (AERA et al., 2014; Messick, 1989).
- *Evidence based on consequences of testing.* In recent years researchers have started examining the consequences of test use, both intended and unintended, as an aspect of validity. In many situations the use of tests is based largely on the assumption that their use will result in some specific benefit (AERA et al., 2014; McFall & Treat, 1999). This line of validity evidence asks the question, “Are these benefits being achieved?”

The overarching goal of your validity studies is to allow you to integrate validity evidence from multiple sources into a sound validity argument (AERA et al., 2014). This suggests that you should pursue multiple studies examining different sources of validity evidence. While it is important to examine multiple sources of validity evidence, the proposed interpretation of your test scores will make specific sources of validity evidence more relevant than other sources. For example, with tests of academic achievement, validity evidence based on test content has typically been viewed as central and essential. If the intended use of your test scores is predicting performance on a criterion, then test-criterion studies examining the ability to predict the criterion should be given the highest priority.

While we have emphasized a number of distinct approaches to collecting evidence to support the validity of score interpretations, validity evidence is actually broader than the strategies described here might imply. The Standards (AERA et al., 2014) state:

*Ultimately, the validity of an intended interpretation of test scores relies on all the evidence relevant to the technical quality of a testing system . . . and include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling; equating, and standard setting; and careful attention to fairness for all examinees. (p. 22)*

**Table 2.5** Phase IV—plan implementation

- |  |
|--|
| Step 1: Submit test proposal if publication is desired |
| Step 2: Implement plan, reevaluate, and modify test    |
| Step 3: Prepare manual                                 |

In other words, when considering the validity of score interpretations, you should consider in total the evidence of the technical quality of the test. The five sources of validity evidence described here are central to building a validity argument, but other information should be carefully considered. Does the test produce reliable scores, is the standardization sample representative, and is there adequate standardization of both administration and scoring? In sum, is the test a well-developed and technically sound instrument? (Table 2.5).

## Phase IV: Plan implementation

The final phase of test development is actually implementing the test plan. An important decision to make is whether you will pursue publication of your test. As we noted earlier, test development is almost always a team effort, and having the backing and support of a test publisher can be extremely helpful, especially if your test plan involves large scale data collection (e.g., nationally representative normative or reference group). We do recognize that many tests are developed for specific research or other applications that do not require large scale data collection and can be completed by a small group of researchers.

### ***Submit a test proposal***

If you are interested in publishing your test you will need to submit a test proposal to a publisher. Test publishers typically encourage authors to contact them fairly early in the process. At the same time, you will need a comprehensive well-developed plan, and ideally some pilot data, to submit to a potential publisher. Robertson (2003) provides an excellent discussion of the test publication process, beginning with the publisher evaluating the test proposal through developing a marketing plan and launching the publication. Major test publishers have slightly different guidelines for submitting a test proposal and these are typically available on their website.

### ***Implement plan, reevaluate, and modify test***

Flexibility is essential in implementing your plan. You will likely discover that some of your “favorite” items have poor item statistics and do not make the final cut. You might find that your instructions for administering, taking, and scoring the test are not clear and need revising. It is possible that the results of your factor

analyses will not be consistent with your hypothesized structure and you need to modify the overall structure of the test. Plan implementation requires that you be data driven and willing to reevaluate and modify your plan as new information becomes available. Never be reluctant to make changes dictated by the data or feedback from those administering the test and collecting the data.

### ***Prepare the test manual***

Once the developmental process is complete you can start finalizing the test manual. The test manual should describe the test development process in sufficient detail that researchers can replicate the process and users can make informed decisions about the usefulness of the test for their applications. Parts of the manual such as administration and scoring instructions should now be in their final form. In the manual you want to introduce and describe the test and its theoretical and conceptual foundations, describe the interpretive schema, and include chapters on standardization, reliability/precision, and validity. The *Standards* ([AERA et al., 2014](#)) provide detailed information on the information that should be included in a test manual, including a chapter outlining the standards for supporting documentation for tests.

## **Concluding comments**

The chapter has emphasized the development of a strategic plan for test development as opposed to a step-by-step description of implementing the plan. For a practical guide to the day-to-day aspects of carrying out such a plan, we recommend a chapter by [Robertson \(2003\)](#), who was for many years in charge of developing and carrying out such plans for several major publishers and who generously has shared his expertise and experiences with the profession. We also recommend anyone considering developing a test acquire and read thoroughly the *Standards* ([AERA et al., 2014](#)) as they provide invaluable guidance to the aspiring test author.

## **References**

- Aiken, L. R. (2000). *Psychological testing and assessment*. Boston: Allyn & Bacon.
- American Educational Research Association, American Psychological Association, and National council on Measurement in Education (2014). *Standards for educational and psychological measurement*. Washington, DC: American Educational Research Association.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook. Cognitive domain*. White Plains, NY: Longman.

- Bush, S., Ruff, R., Troster, A., Barth, J., Koffler, S., Pliskin, N., ... Silver, C. (2005). Symptom validity assessment: Practice issues and medical necessity. *Archives of Clinical Neuropsychology, 20*, 419–426.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
- Cronbach, L., Rajaratnam, N., & Gleser, G. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.
- Embreton, S., & Reise, S. (2000). *Item response theory for psychologists*. London: Taylor & Francis.
- Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule: I. Construction of the schedule. *Journal of Psychology, 10*, 249–254.
- Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory* (Rev. ed). Minneapolis: University of Minnesota Press.
- Hays, W. (1994). *Statistics* (5th ed). New York: Harcourt Press.
- Hopkins, K. D. (1988). *Educational and psychological measurement and evaluation* (8th ed). Boston: Allyn & Bacon.
- Kelly, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17–24.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of reliability. *Psychometrika, 2*, 151–160.
- Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika, 17*, 181–194.
- McFall, R. M., & Treat, T. T. (1999). Quantifying the information value of clinical assessment with signal detection theory. *Annual Review of Psychology, 50*, 215–241.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Upper Saddle River, NJ: Merrill Prentice Hall.
- Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior assessment system for children*. Pearson Clinical Assessment: Bloomington, MN.
- Reynolds, C. R., & Livingston, R. B. (2012). *Mastering modern psychological testing: Theory & methods*. Boston: Pearson.
- Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed). Boston: Allyn & Bacon.
- Robertson. (2003). A practical model for test development. In C. R. Reynolds, & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Vol. I. Intelligence, aptitude, and achievement* (2nd ed., pp. 24–57). New York: Guilford Press.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.
- Woodcock, R. W. (1999). What can Rasch-based scores convey about a person's test performance? In S. E. Embretson, & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 105–127). Mahwah, NJ: Erlbaum.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III (WJ III) Complete Batter*. Itasca, IL: Riverside.

## **Part III**

# **Assessment of Intelligence**

# Interpreting pediatric intelligence tests: a framework from evidence-based medicine

3

Andrew J. Freeman and Yen-Ling Chen  
University of Nevada, Las Vegas, NV, United States

## Introduction

Formal assessment of children and adolescents via intelligence tests for educational purposes is now over 100 years old. In the United States, approximately 13% of school children receive special education services ([U.S. Department of Education, 2016](#)). The administration and interpretation of an intelligence test likely plays a role in the decision-making process. However, substantial fault lines regarding the theory of intelligence, utility of intelligence testing, and interpretations of intelligence tests remain. Major theorists do not fully agree on the definition of intelligence. A consensus definition from psychologists studying intelligence is that intelligence represents a general mental capability in which “individuals differ from one another in their ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, and to overcome obstacles by taking thought” ([Neisser et al., 1996](#)). The consensus definition accounts for both the interrelated nature of cognitive abilities (i.e., shared variance, or  $g$ ) as well as potentially distinct systems existing in intelligence (i.e., specific abilities). More recently in the context of artificial intelligence, computer scientists defined intelligence by integrating multiple definitions of human intelligence and proposed that an overarching definition of intelligence is “an agent’s ability to achieve goals in a wide range of environments” ([Legg & Hutter, 2007](#)). The computer science definition of intelligence explicitly suggests a general overall ability, but implicitly suggests that intelligence might consist of specific abilities as well. In both psychology and artificial intelligence, the definitions of intelligence note that an individual’s performance on measures of intelligence will vary across the specific domains measured and may vary when judged by different criteria or at different times. How much the differences observed across domains and intelligence tests matter in clinical settings is a subject of great debate. Therefore, the purpose of the current chapter is to lay a foundation for the interpretation of intelligence tests with children and adolescents. The chapter is structured in the following format: (1) a brief history of intelligence testing focusing on the content of clinical

intelligence tests, (2) a framework for considering different interpretation approaches to intelligence tests from an evidence-based medicine perspective, and (3) a brief review of two current intelligence tests commonly used with children and adolescents.

## A brief history

Intelligence as a concept is substantially older than formal intelligence testing. Philosophers espoused versions of intelligence theory from at least 300–400 BCE. For example, Plato and Aristotle discussed the concept of intelligence (Zusne & Zusne, 1984). Plato believed that the soul was divided into three components: reason, will, and appetite. The purpose of reason was to regulate and control the body and its desires (Robinson, 1972). Aristotle, Plato's student, taught that excellence in logical and semantic operations were critical to higher order thinking. Additionally, Aristotle believed that individuals must possess a minimum amount of critical thinking ability to engage in philosophical activities (Tigner & Tigner, 2000). In particular, Aristotle proposed two types of mental processing—cybernetic and orectic. Cybernetic processing relates to thought processes, whereas orectic processing defines moral and emotional reasoning (Das, Cummins, Kirby, & Jarman, 1979). Continuing in the philosophical tradition, philosophers such as Aquinas, Kant, and Hobbes all espoused views on intelligence. However, most philosophers did not attempt to directly assess intelligence. In the 16th century, Fitzherbert and Huarte stated that an individual's intelligence could be tested (Sattler, 2008). A common thread across the early philosophers is that intelligence is a general construct. Despite these early contributions to the construct of intelligence, the formal, modern study of intelligence and intelligence testing had its foundations in 19th century Europe.

In the early 1800s, French physicians distinguished between mental health and physical health. In 1828, Jean Esquirol separated mental health into two types. Esquirol proposed that there were “mentally deranged persons”—individuals with severe mental illness—and “idiots”—individuals with intellectual disability (Esquirol, 2012). This separation continues in the modern day in the form of distinct funding streams and services for physical illness, mental illness, and intellectual disability. Esquirol examined the language and verbal characteristics of individuals with intellectual disabilities and proposed a classification system based on severity of speech impairments. Esquirol's system considered the most severely impaired to have no speech, and as language abilities improved (e.g., the use of single syllables or short words or phrases), so did intellectual abilities. Esquirol focused on verbal abilities as a general marker of intellectual ability. In contrast, Édouard Séguin, another French psychiatrist, focused on performance abilities through examination of sensory discrimination and motor control to classify individuals with intellectual disabilities by severity. For example, Séguin used a pegboard-based motor performance task similar to the modern Tactual Performance

Test ([Boake, 2002](#)). Séguin believed that performance on motor tasks was a valid marker of an individual's intellectual capacity. Both Esquirol and Séguin developed methods for measuring intelligence and systems for classifying individuals with intellectual disability. Additionally, these two pioneers provided a framework for considering intelligence along a gradient of ability that could be measured via verbal or performance methods.

In 1869, Francis Galton published *Hereditary Genius*. Galton's book reflected data collection of what he considered intellectual ability across social strata of England. His book attempted to explain why "genius" appears in family trees by both measuring intelligence and applying the theory of evolution to intelligence. In doing so, Galton popularized the term intelligence from the Latin verb *intelligere* which means to comprehend or perceive. Francis Galton's work focused on the systematic measurement of psychological and physical processes that he believed to be associated with intelligence. However, he focused on poor measures of intelligence such as sensation (e.g., sensory acuity), perception (e.g., the ability to discriminate pitch), and physical measurements (e.g., height, weight). Meanwhile in Germany, Wilhelm Wundt built the first psychology laboratory to systematically measure mental processes. Many of Wundt's students contributed to both intelligence theory and intelligence testing. For example, Emil Kraepelin (also known as the father of the DSM) devised tests to measure perception, memory, and attention. At the close of the 19th century, Galton and Wundt's work strongly influenced the idea that intelligence is a psychological construct that could be systematically measured.

James McKeen Cattell, a student of both Galton and Wundt, founded a psychology laboratory at the University of Pennsylvania. Cattell examined intelligence using a blend of Galton and Wundt's methods resulting in a weak measure of general cognitive ability. His most important contributions to intelligence testing were: (1) translating earlier philosophical work into applied measures, (2) standardizing assessment methods, and (3) training the first generation of American psychologists ([Sternberg, 1990](#)). As Cattell was helping to found psychology as a discipline in the United States, French psychologists developed the first modern intelligence test primarily focusing on verbal abilities. In France, the education minister requested a practical method for determining which students were intellectually disabled and could be excluded from regular education. In response to this request, Alfred Binet, Victor Henri, and Theodore Simon transitioned intelligence tests from basic evaluations of sensory processing to more complex mental processes—such as language, learning and memory, judgment, and problem-solving—that could be used to classify students based on cognitive ability. The 1905 Binet–Simon Scale—the prototype for future intelligence tests—measured intelligence by assessing verbal ability on relatively novel items across many domains (e.g., comprehension questions that measured a child's response to an abstract question, digit span via repetition of numbers, and vocabulary measured via definitions of concrete terms). Scores on the Binet–Simon Scale were determined relative to an individual's chronological age, and were used to make a single determination of cognitive ability. The Binet–Simon Scale is the

forerunner of most modern intelligence tests in that it assesses a general ability through multiple lower order mental processes ([Humphreys, 1994](#)).

In 1908, Henry H. Goddard translated the Binet–Simon Scale into English and implemented it as a testing device at the Vineland Training School in New Jersey. Goddard's implementation followed minor revisions to the scale as well as a larger standardization sample of 2000 American school-age children. Similar to Binet, Goddard's modifications resulted in measuring intelligence from a variety of abilities and using the test to inform a single decision of whether a child was intellectually impaired. In 1916, Lewis Terman released the Stanford Revision and Extension of the Binet–Simon Scale, also known as the Stanford–Binet Intelligence Scale. The Stanford–Binet followed more modern psychometric practices in that it was normed based on the performance of 1000 school-age children (4–14 years) from a more representative sample of middle-income Californian communities. The Stanford–Binet also introduced the intelligence quotient (IQ) which was initially the ratio between mental age and chronological age ( $\text{IQ} = (\text{mental age}/\text{chronological age}) \times 100$ ). The IQ score resulted in a focus on  $g$ , the general factor of intelligence proposed by [Spearman \(1904\)](#). However, the Binet–Simon and Stanford–Binet continued to focus primarily on verbal abilities as a measure of intelligence. By 1916, many of the core principles seen in current intelligence tests were developed. However, individual tests tended to focus either on verbal performance or nonverbal performance (e.g., the Army alpha as a measure of Verbal Intelligence and Army Beta as a measure of Performance Intelligence) and items were scored as correct or incorrect. Therefore, the critical development of combining verbal measures and nonverbal measures of intelligence into a single test and potentially scoring items on a rank-order scale were still required.

Scoring of intelligence tests was primarily defined by correct/incorrect responses. These scores were summed to a total and then used to estimate a mental age equivalent. Robert Yerkes believed that scores could reflect more information than simply correct/incorrect responses. Yerkes introduced the modern point-scale format in which questions could be scored via gradients in quality of correctness (e.g., 0 = incorrect, 1 = correct, 2 = ideal) and/or speed to completion. Point-scoring underpinned revisions to the Stanford–Binet in 1937. By the 1930s, most intelligence tests primarily focused on providing a general estimate of intellectual ability via measuring either verbal or performance abilities. In 1939, David Wechsler developed the Wechsler–Bellevue Scale. The first Wechsler scale heavily influenced future intelligence test development and interpretation in three ways. First, the Wechsler test introduced not only the full scale IQ (a measure of  $g$ ) but also the verbal IQ and performance IQ. Wechsler's test returned to the earlier roots of intelligence theory of Esquirol and Seguin ([Wechsler, 1939](#)) and unified the measurement into a single measure by including tests of verbal ability and performance ability. Second, Wechsler introduced norm-based scoring instead of age-based IQ ratios. Scores on Wechsler's test represented a person's ability relative to peers and not a person's mental age relative to one's chronological age ([Thorndike & Lohman, 1989](#)). Third, Wechsler heavily borrowed from existing intelligence tests

to create a measure using what he considered to be the best available measures of intelligence (Sattler, 2008). In 1949, Wechsler published the Wechsler Intelligence Scales for Children (WISC), introducing these concepts to intelligence testing of children and adolescents. Taken together, the fundamental underpinnings of modern intelligence tests are: (1) a standardized measure to classify individuals on general cognitive ability, (2) general cognitive ability measured across a variety of inter-related domains and methods, (3) general cognitive ability measured through relatively novel or unique tasks, (4) point-system scoring at least for some subtests, and (5) scores that reflect an individual's ability relative to same-age peers. While many versions and editions of intelligence tests have been developed since the 1940s, these core principles underlie modern intelligence tests regardless of publisher and author.

Numerous intelligence tests have been developed since the 1940s. Some tests are primarily revisions and updates to prior versions (e.g., Wechsler tests), while others offer more dramatic changes to methods of intelligence testing (e.g., Stanford–Binet, Woodcock–Johnson). Recent revisions to major intelligence tests generally represent only minor changes (e.g., updated norming samples, revised item sets) and not wholesale changes in theory or approach to intelligence tests (Naglieri, 2000). As intelligence tests evolved to include more domains assessed and engendered more widespread use, theories of intelligence increased in complexity. From the earliest theories of intelligence, debate concerning whether intelligence was a single construct (e.g., Spearman, 1904), an interrelated set of abilities that were difficult to distinguish (e.g., Binet & Simon, 2003), or specific mental abilities that were correlated (e.g., Thurstone, 1934) has existed. Early tests (e.g., Army alpha, Army beta, Binet–Simon) viewed results of testing as reflective of underlying  $g$ —general ability. By the 1930s, some theorists suggested that intelligence consists of primary abilities and that  $g$  may not exist (e.g., Thurstone, 1934). These early theories have continued to evolve. Modern structural theories of intelligence tend to incorporate  $g$  as either a hierarchical ability that directly influences broad abilities (e.g., Schneider & McGrew, 2012), or a general ability that directly influences task ability in addition to other cognitive abilities (Carroll, 1993). Since the 1990s, commercial intelligence tests began to explicitly incorporate the theoretical view that tests can measure more than just  $g$  (Frazier & Youngstrom, 2007). However, the degree to which structural theories of intelligence have been incorporated into clinical assessments is debatable. Additionally, whether theories of intelligence can be directly applied to clinical utility is an empirical question. Therefore, the interpretation of modern intelligence tests should weigh available empirical evidence more strongly than theoretical considerations.

Fundamental to the debate on how to interpret intelligence tests is the weight to which  $g$  should be given. Substantial theoretical debate exists on what  $g$  represents. Is  $g$  a higher-order factor that directly influences lower order abilities (McGrew, 2009), an independent bifactor (Beaujean, 2015), or simply a formative construct representing the positive manifold of abilities that should not be strongly interpreted (van der Maas, Kan, & Borsboom, 2014)? Current evidence, as measured on modern intelligence tests, does not provide an overwhelmingly clear answer to this

question. However, most major theorists appear to agree that at the minimum,  $g$  is measured and represented on clinical intelligence tests. The debated question is whether clinical intelligence tests measure more than  $g$ . Most germane to pediatric psychologists is whether modern intelligence tests measure specific abilities such as executive functioning, attention, or processing speed in addition to a more general common factor such as  $g$ .

## **Framework for interpreting intelligence tests**

As modern intelligence tests developed, so did the debate on how to interpret the results of the tests. Interpretation of intelligence tests consists at the most basic level of the assessor choosing the relative weight to place on two dimensions: qualitative–quantitative and idiographic–nomothetic (Kamphaus, Winsor, Rowe, & Kim, 2012; Kranzler & Floyd, 2013). First, an assessor must determine how heavily to weigh qualitative or quantitative interpretations. Qualitative interpretation focuses on the process of testing or how an individual completes the test. Quantitative interpretation focuses on the empirical scores of an individual as the summary piece of interpretative information. Second, an assessor must determine how idiographic or nomothetic in interpretation to be. Idiographic interpretation focuses on the uniqueness of the person as identified by within person performance over time or across abilities. Nomothetic interpretation focuses on how a person performs relative to a more general population represented in the test's norm-group. In general, interpretation of children's intelligence test will rely on using all the available blends of the qualitative–quantitative and idiographic–nomothetic dimensions—qualitative-idiographic, qualitative-nomothetic, quantitative-idiographic, and quantitative-nomothetic. However, substantial debate remains in how to weigh the information gathered from the different approaches. Therefore, a framework for the evaluation of evidence from children's intelligence tests is necessary.

Evidence-based medicine (EBM) provides a general framework for the evaluation and management of the information overload that occurs in the context of assessment. EBM focuses on the utility of a test and not the tests theoretical merits. EBM defines the utility of a test as the extent to which testing improves outcomes relative to the current best alternative, which could include either a different test or no testing at all (Bossuyt, Reitsma, Linnet, & Moons, 2012). EBM calls on test interpretation to add value to an individual that outweighs any potential harms such as misdiagnosis, recommendations for ineffective treatments, or the cost of testing (Straus, Glasziou, Richardson, & Haynes, 2010). In the context of psychological assessment, utility can be defined by the “Three Ps” (Youngstrom, 2008, 2013). *Prediction* is the concurrent or prospective association between the test outcome and a criterion of importance such as diagnosis, competence, or lack of effort. One example of a prediction question is: “In children, would low processing speed scores compared to normal processing speed lead to an accurate diagnosis of

ADHD?" *Prescription* refers more narrowly to the test's ability to change the course of treatment or select a specific treatment. An example of a prescription question is: "In children, do children with low processing speed benefit from extended test-taking time compared to normal time limits for completing school exams?" *Process* represents the outcome's ability to inform about progress over the course of treatment and quantify meaningful outcomes. One example of a process question is: "In children, will scores on an intelligence test reflect improvement after a traumatic brain injury?" Applying the 3Ps to children's intelligence test interpretation provides the strongest and most persuasive reasons for completing an assessment (Meehl, 1997) because they directly link assessment to treatment and prognosis. Therefore, each of the four methods for interpreting intelligence tests will be evaluated in the context of the "Three Ps" of assessment.

## Qualitative-idiographic approaches

Qualitative-idiographic approaches focus interpretation on the clinician's observation of a person's ability and/or behavior as it changes over the course of testing or relative to the person's behavior outside of testing. The practice of assessment creates a rich set of observational data that have traditionally been a highly prized gathering of information in neuropsychological assessment (Semrud-Clikeman, Wilkinson, & Wellington, 2005). For example, clock drawing tests are scored on their qualitative merits as to whether they were drawn correctly or incorrectly. The type of mistakes made while an adult draws the clock are associated with dementia and hemispheric neglect (Agrell & Dehlin, 2012). The emphasis on observation of behavior and changes in behavior applies to the assessment of children as well. Most available texts on assessment strongly recommend attending closely to an examinee's behavior during testing (Sattler, 2008). Anecdotal justifications of behavior during testing such as an observant clinician identifying *petit mal* seizures are given as reasons for attending to behavioral observations. However, an observant clinician would likely notice *petit mal* seizures occurring regardless of the testing circumstances as they are not a function of the testing procedure. For rare disorders, diseases, or behaviors, detailed observation of qualitative behavior likely has predictive value. However, placing strong emphasis on qualitative-idiographic observations places the clinician at risk of cognitive biases such as confirmation bias (e.g., "I'm assessing for ADHD, the child is fidgeting, and so this must be ADHD"), use of personal norms that may or may not be accurate, and overemphasis on data from a single method at a single point in time. Therefore, the predictive utility of qualitative-idiographic approaches to intelligence test interpretation is likely low.

Most psychologists assess more than rare or extremely odd presenting problems. Most clinical assessments of children focus on high base rate concerns such as attention-deficit hyperactivity disorders, learning disabilities, and traumatic brain injuries. In this context, the utility of qualitative observations during intelligence

testing is questionable. Neuropsychology has developed alongside advances in medical imaging that has made the use of neuropsychological tests for specific diagnosis less critical. Instead, neuropsychological assessment has focused on how results inform process. From a qualitative-idiographic perspective, a clinician might examine the process of how a person completes testing. For example, an examiner might notice that a child smiles during some tests and not others. One could interpret these qualitative observations of idiographic change in smiling to indicate genuine happiness with their performance on specific tests. However, smiling can indicate several other emotions other than happiness, such as fear and contempt (Ekman, 1984), or smiling can be used interpersonally—such as smiling to indicate compliance, cooperation, or that one is listening (Fridlund, 1991). Therefore, idiographic-qualitative interpretations of test session behavior require that the examiner must decode the meaning of the behavior based on limited information that is at high risk for cognitive biases.

In addition to difficulties in identifying the reason for a qualitative-idiographic observation during testing, the observation should also display prescriptive utility. Based upon observations during testing, how might one intervene in another setting? Test recommendations are often based on this approach. For example, if a specific one-on-one intervention works during testing (e.g., frequent breaks), then the clinician might recommend the same intervention to other settings (e.g., frequent breaks while completing homework). While coherent in thinking, the reality is that test session behaviors are only weakly correlated with other behaviors in other settings (McConaughy, 2005). More importantly, most recommendations for how to write treatment recommendations include cautions against making broad recommendations based on a limited sample of behavior (e.g., Sattler, 2008). Therefore, the prescriptive utility of qualitative-idiographic interpretations is likely low.

In contrast to weaknesses in highlighted external validity, the qualitative-idiographic observations may have *predictive* utility in terms of the validity of a test administration. For example, an examiner might observe changes in a child's cooperativeness over the course of testing and hence integrate that information into the judgment of the validity of that test or weight placed on the child's performance on that specific test. In more systematic evaluations of test session behavior, factors such as cooperativeness, attention, and avoidance are associated with test performance (Daleiden, Drabman, & Benton, 2002; McConaughy, 2005; Oakland, Callueng, & Harris, 2012). Therefore, the utility of qualitative-idiographic interpretations of a child's performance during or on an intelligence test is likely most useful for evaluating whether the test administration should be interpreted as valid.

## **Qualitative-nomothetic approaches**

Qualitative-nomothetic approaches are rare in practice. Like the more idiographic approaches, the qualitative-nomothetic approach is founded in the idea that careful,

systematic observation of a child's process for solving a problem yields useful information about that child's cognitive functioning separate from the final solution (Kaplan, 1988; Werner, 1937). What separates this approach from the qualitative-idiographic approach is that qualitative-nomothetic approaches score a person's process of test-taking and those scores are compared to normative data. Qualitative-nomothetic approaches can be separated into major categories: (1) normative evaluations of process during testing, and (2) systematic behavioral observations. The integrated versions of the Wechsler Intelligence Scales for Children (Wechsler & Kaplan, 2015) are examples of the normative evaluations of process during testing approach. From a behavioral perspective, standardized measures such as the Guide to the Assessment of Test Session Behavior (Glutting, Oakland, & Konold, 1994) or the Test Observation Form (McConaughy & Achenbach, 2004) provide a metric for interpreting a child's test performance relative to same-age peers. However, normative interpretations of qualitative observations are rare in the empirical literature, particularly for qualitative-normative interpretations of test-taking processes. Therefore, the interpretation of these procedures from the 3P perspective is limited (Hebgen & Milberg, 2009; Poreh, 2006).

Evaluations of the process of completing an intelligence test rests on the core principle that subtests of intelligence tests are not pure measures of a specific neurocognitive process but instead represent the combination of many basic neurocognitive functions. For example, responding accurately to a vocabulary item involves multiple cognitive processes such as the ability to understand directions and stimulus words (i.e., receptive language), knowledge of the word's definition (i.e., semantic knowledge), the ability to retrieve information from long-term memory, and the ability to verbally respond to the word (i.e., expressive language) as well as many other processes (Sattler, 2008). While one cannot remove these processes from higher-order tasks, one can change the relative importance of these distinct processes by changing the modality of administering the items. For example, the WISC Integrated Vocabulary subtest removes free-response as the answer modality and replaces it with multiple choice responses. In doing so, the child's process of answering the question may rely less on memory as the correct answer is present and does not need to be as strongly recalled, and also the response may rely less on expressive language as the child can point to the correct answer. The difference between free-recall and multiple choice recall are compared, as well as behavioral observations about the answering process. In practice, these observations are then compared to theoretical conceptualizations of processing for that ability (McCloskey & Maerlender, 2005). However, current empirical evidence is lacking for the utility of this approach. In theory, the process approach has the potential to provide clinically useful information regarding *prediction*, *process*, and *prescription*. If the observed differences in processing are both sensitive and specific to particular outcomes, then the qualitative-nomothetic approach may have *predictive* utility. If the observed processing differences are sensitive to change over time, then the qualitative-nomothetic approach may have *process* utility. If the observed processing differences can be remediated by specific interventions, then the qualitative-nomothetic approach may have *prescriptive* utility. In summary, while theoretically

promising, the qualitative-nomothetic interpretation of intelligence tests has limited empirical support.

In contrast to process models of intelligence tests, the systematic observation of test session behavior may have some utility. Like process measures during intelligence tests, the evidence-base for test session observations is relatively small. The use of test session behavior to *prescribe* treatments is unknown. From a *process* perspective, if the outcome of interest was improvement during test sessions, then the measures could function as a process measure. However, test session observations are often used to make *predictions* about more general functioning. Using test session behavior for prediction is problematic because the behavior of children and adolescents tends to be situationally specific (De Los Reyes et al., 2015). Like cross-informant measures, the correlation between test session behavior and exosession behavior is low and suggestive of low *predictive* utility (Daleiden et al., 2002; McConaughy, 2005).

In the context of determining whether the youth's behavior resulted in valid test scores, the *predictive* utility of test session observations is substantially higher. Poor test session behavior (e.g., higher levels of uncooperativeness, inattention, and avoidance) are correlated ( $r_s$  = approximately  $-.3$ ) with poorer performance on intelligence tests (Glutting, Youngstrom, Oakland, & Watkins, 1996). Clinicians must be careful in interpreting the relationship. A negative correlation could occur because the youth has a disorder that causes lower IQ and/or difficulties during testing, or the process of testing is difficult and causes problematic behaviors to occur. Regardless of the cause, poor test session behavior is associated with lower performance on intelligence tests. In cases of poor test session behavior, clinicians should carefully consider whether the results of testing represent a valid evaluation.

## Quantitative-idiographic approaches

Quantitative-idiographic approaches are common in practice among neuropsychologists, child psychologists, and school psychologists. The quantitative-idiographic approach is founded on the idea that summary scores such as an overall IQ score mask individual variations in a child's neurocognitive ability (Lezak, 1988). Children with different illnesses, disorders, and disability-related backgrounds tend to show distinct profiles relative to healthy controls on specific neurocognitive abilities. For example, children with ADHD tend to display a profile of lower scores on overall intelligence, attention, executive functioning, and working memory relative to healthy controls (Frazier, Demaree, & Youngstrom, 2004; Pievsky & McGrath, 2018). In contrast, youth with pediatric brain tumors tend to perform lower on overall IQ, performance IQ, and verbal IQ relative to healthy controls (De Ruiter, Van Mourik, Schouten-Van Meeteren, Grootenhuis, & Oosterlaan, 2013). These differences in profiles have led many to propose different systems for comparing a youth's profile on intelligence tests to make predictions about diagnosis, monitor

process, and prescribe interventions (cf. Fiorello, Hale, & Wycoff, 2012; Fuchs, Hale, & Kearns, 2011; Kaufman, Raiford, & Coalson, 2016; McGrew, 2009; Sattler, 2008). Therefore, consideration to the empirical evidence for ipsative-based interpretation approaches is critical.

Different systems for conducting ipsative comparisons exist. For example, some focus on the neuropsychological functioning underpinning specific subtests (Fiorello et al., 2012), while others focus on structural theories of intelligence such as CHC Theory (Kaufman et al., 2016; McGrew, 2009). As a result, specific interpretations of ipsative approaches can lead to differing conclusions. However, the ipsative interpretation systems share a commonality in process. First, neuropsychologists are instructed to start by comparing the components or factors of intelligence tests (e.g., processing speed compared to verbal comprehension on the WISC-5). Second, neuropsychologists are encouraged to compare individual subtests to each other to make interpretations about specific abilities. By doing this, neuropsychologists should be able to make individualized hypotheses for why behavior is occurring based on a nonexhaustive, selective sampling of mental functioning under fixed conditions (Kaufman & Lichtenberger, 2006). The tenets and results of this approach are intuitively appealing and match many of the general recommendations for neuropsychological assessment, such as comparing neurocognitive domains to each other (Donders, 2016; Postal et al., 2018; Tzotzoli, 2012). However, applying an ipsative approach in order to generate hypotheses and treatment recommendations from an intelligence test is most likely empirically incorrect for a variety of reasons, including technical and practical reasons. Therefore, a brief review of the empirical evidence for ipsative interpretation follows.

Early recommendations from proponents of ipsative approaches to interpreting childhood intelligence tests focused on the ability to predict clinically meaningful diagnoses such as learning disabilities, ADHD, and brain injury. For example, youth with ADHD are hypothesized to show the ACID profile. The ACID profile consists of poor performance on Arithmetic, Coding, Information, and Digit Span subtests of the WISC. Unfortunately, the ACID profile is a poor predictor of ADHD diagnosis (Ek et al., 2007). In fact, many of the hypothesized ipsative differences are poor predictors of their targeted outcomes. Profile differences on intelligence tests do not predict reading difficulties (Stuebing et al., 2002), math difficulties (Hale, Fiorello, Bertin, & Sherman, 2003), behavioral outcomes (Oh, Glutting, & McDermott, 1999), special education placement (Glutting, McDermott, Konold, Snelbaker, & Watkins, 1998; Glutting, McDermott, Watkins, & Kush, 1997), or academic achievement (McDermott, Fantuzzo, & Glutting, 1990; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). The ipsative approaches tend to have relatively high specificity but very low sensitivity (Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016). Consider the examples of youth with unilateral brain damage. In individuals with unilateral left hemisphere brain injury, Verbal IQ tends to be lower than Performance IQ. In individuals with unilateral right hemisphere brain injury, Performance IQ tends to be lower than Verbal IQ (Kaufman & Lichtenberger, 2006). Knowing that a youth has unilateral brain injury then conducting an assessment that verifies this finding

is one approach to confirming or explaining brain injury. It is substantially more difficult to infer brain injury from this profile on clinical intelligence tests because many diseases, disorders, and disabilities could give a similar core profile. The sensitivity of the profile is low because not all youth with traumatic brain injury will have this profile of impairment. Empirical evidence strongly indicates that the effectiveness of ipsative interpretation of intelligence tests for the purpose of predicting an individual diagnosis should be considered a myth (Macmann & Barnett, 1997; Watkins, 2003). Some proponents of this approach have even begun to caution against making strong interpretations based on ipsative interpretations (Flanagan & Alfonso, 2017). Therefore, interpreting profiles of childhood intelligence tests for the purpose of predictions is generally not recommended until consistent positive evidence of predictive utility is published.

In contrast to predicting a specific outcome (e.g., diagnosis) via ipsative interpretation, a neuropsychologist might want to monitor the process of treatment for youth with a specific disorder, illness, or disability. For example, do the results of intelligence tests allow for the monitoring of recovery from traumatic brain injury? The precise answer to this question depends upon how one interprets change. Ipsative interpretations of intelligence tests are not recommended for predictive purposes because of low sensitivity to deficits. In longitudinal studies, ipsative interpretations are highly unstable and tend to remain stable at levels only slightly better than chance (Livingston, Jennings, Reynolds, & Gray, 2003; McDermott et al., 1992; Watkins & Canivez, 2004). However, monitoring treatment outcomes via intelligence tests and other neuropsychological tests is critical. Intelligence tests, and their component pieces, tend to be very stable over time much like neuropsychological tests (Calamia, Markon, & Tranel, 2013). Therefore, changes in scores on intelligence tests, their components, and possibly their subtests could be used to monitor the process of treatment and recovery.

Quantitative-idiographic assessment of the outcomes of treatment is relatively common in neuropsychological research. The focus of interpretation is on monitoring change in a specific neuropsychological domain across testing sessions and not the blending or contrasting of domains as in the ipsative approaches. For example, longitudinal monitoring of youth with traumatic brain injury indicates that youth with mild traumatic brain injury recover more rapidly and at higher rates than adults with mild traumatic brain injury (Königs, Engenhorst, & Oosterlaan, 2016). On the negative side, process monitoring also reveals that youth with severe traumatic brain injury tend to have slower rates of recovery than adults with severe traumatic brain injury (Königs et al., 2016); youth born prematurely tend to have long-term deficits in overall cognitive ability, executive functioning, and processing speed that tend to be stable (Brydges et al., 2018); and that some procedures such as revisions of shunt placements in youth with hydrocephaly may lower cognitive performance (Arrington et al., 2016). However, group level changes over time or from treatment are not necessarily clinically informative to an individual (Jacobson & Truax, 1991). In an individual youth's case, a neuropsychologist wants to answer the question of, "Is the change on this child's intelligence test indicative of actual change?" To answer this question, Jacobson and Truax (1991) proposed the creation

of a reliable change index (RCI) that accounts for fluctuations in scores due to measurement error.

Shunt placement revisions for youth with hydrocephalus, excessive accumulation of fluid in the brain, tends to lower overall cognitive abilities by approximately 3 points (Arrington et al., 2016). Imagine examining the same child pre- and post-shunt revision. Say, for example, the child's FSIQ has changed by 3 points on the WISC-V. Has this child's procedure significantly changed their intellectual ability? Jacobson and Truax's RCI represents the amount of change necessary to consider the change to be due to more than measurement error. The RCI is calculated as follows:

$$\text{RCI} = \frac{T_2 - T_1}{S_{diff}}.$$

$S_{diff}$  represents the standard error of the difference between two scores which is the spread of change scores that would be expected if no change occurred. The  $S_{diff}$  is calculated by:

$$S_{diff} = \sqrt{2(S_1 \sqrt{1 - r_{xx}})^2}.$$

$S_1$  is the standard deviation of the measure.  $r_{xx}$  is the test-retest reliability of the measure. The RCI is then compared to the  $z$ -score that matches how confident one would be in the change not occurring due to chance (e.g., 95% confidence would use a  $z = 1.96$ , 90% uses a  $z = 1.64$ ). In our example of the child, the standard deviation of the FSIQ on the WISC-V is 15 and the test-retest reliability is .92. A rearrangement of the formula indicates that for 95% confidence ( $z = 1.96$ ),  $\Delta\text{FSIQ} = 1.96 \times S_{diff} = 1.96 \times \sqrt{2((15\sqrt{1 - .92})^2)} = 11.76$ . So, for youth whose FSIQ changes by 12 points or more one would conclude that the youth's score has reliably changed. See Table 3.1 for reliable change calculations based on Jacobson and Truax (1991) for the WISC-5. In our example case of a child's FSIQ changing by 3 points, one would conclude that the child's FSIQ on the WISC-V has not reliably changed.

Jacobson and Truax's RCI has been criticized in both neuropsychology and clinical psychology. Criticisms can generally be considered as either criticisms of the outcome (Kazdin, 2001; Wise, 2004) or criticisms of the mathematical model (Hinton-Bayre, 2016; Temkin, Heaton, Grant, & Dikmen, 1999). Given the scope of practice in neuropsychology, neuropsychologists tend to accept change in neurocognitive abilities as an acceptable outcome measure. Neuropsychological critiques of the RCI focus on the underlying mathematical model, specifically the calculation of the correct error term. Many have proposed adjustments to correct for repeated testing and differences in distributions across time points (Hinton-Bayre, 2016). However, the adjusted RCIs proposed by various authors are more complicated than Jacobson and Truax's RCI and result in relatively minor differences in the

**Table 3.1** WISC-V subscales and indices

Subtest	Typically reported						Optional							
	FSIQ	VCI	VSI	FRI	WMI	PSI	QRI	AWMI	NVI	GAI	CPI	NSI	STI	SRI
Similarities	P	P								P				
Vocabulary	P	P								P				
Information	S	S												
Comprehension	S	S												
Block Design	P		P							P				
Visual Puzzles	S		P							P				
Matrix Reasoning	P			P						P				
Figure Weights	P			P						P				
Picture Concepts	S			S						P				
Arithmetic	S			S										
Digit Span	P				P			P			P			
Picture Span	S				P			P			P			
Letter–Number Sequencing	S				S			P			P			
Coding	P					P			P		P			
Symbol Search	S					P					P			
Cancellation	S					S						P		
Naming Speed Literacy												P		P
Naming Speed Quantity												P		P
Immediate Symbol Translation												P		P
Delayed Symbol Translation												P		P
Recognition Symbol Translation												P		P

P, primary; S, secondary.

amount of change necessary for an individual to display reliable change (Maassen, 2004; Temkin et al., 1999). Additionally, Jacobson and Truax's (1991) RCI formula does not appear to display consistent bias relative to the adjusted RCIs (i.e., over- or underestimating reliable change). Therefore, clinicians monitoring change at the individual level should keep it simple and consider using Jacobson and Truax's RCI to monitor the *process* of treatment or course in their patients.

In many neuropsychology assessments, the purpose of the assessment is to make treatment recommendations. Recommendations are answers to the question: "Do these specific results *prescribe* a particular treatment?" Many scholars and practitioners have argued that cognitive assessments should result in better intervention planning (Feifer, 2008; Fiorello, Hale, & Snyder, 2006). Altering an intervention to account for an individual's neuropsychological profile is conceptually reasonable. Alterations to treatments based on neuropsychological profiles has been widely tested in school-based interventions for learning disabilities. Meta-analysis comparing learning interventions with and without neurocognitive elements indicate little to no difference in outcome (Burns et al., 2016; Stuebing et al., 2015). Interventions targeting the specific skill deficit (e.g., math concepts) account for most of the variance in outcome. Interventions that directly target neurocognitive processes (e.g., working memory training) do not typically have far-transfer effects to outcomes of interest such as academic performance (Sala & Gobet, 2017; Weicker, Villringer, & Thöne-Otto, 2016). Therefore, using children's intelligence test scores to *prescribe* specific interventions for youth is generally not recommended at this time.

## Quantitative-nomothetic approaches

Quantitative-nomothetic approaches are also relatively common in clinical practice. The quantitative-nomothetic approach consists of examining standardized scores on measures and making decisions based on those scores. The quantitative-nomothetic approach differs from the quantitative-idiographic approach in that comparisons within individuals are typically not made. Practitioners that ascribe to the quantitative-nomothetic approach tend not to compare within person abilities (i.e., working memory relative to processing speed). Instead, the quantitative-nomothetic approach focuses on an individual child's relative placement to peers. Test scores for cognitive abilities tend to distribute normally in the general population, so the quantitative-nomothetic approach focuses on a child's relative placement on the normal curve. A specific child's individual difference from peers represents the child's abilities. In the quantitative-nomothetic tradition, the resulting standardized scores are then interpreted as indicative of a child's intellectual functioning.

Most research on the *predictive* utility of intelligence test focuses on the quantitative-normative approach to interpretation. For example, most case-control studies in neuropsychology compare standardized scores between youth with a condition and healthy youth. The use of standardized scores makes the comparisons

quantitative and nomothetic. For example, youth with a history of maltreatment tend to have lower scores on overall cognitive abilities, working memory index, processing speed and attention difficulties relative to youth without a history of maltreatment (Malarbi, Abu-Rayya, Muscara, & Staggatt, 2017; Masson, Bussières, East-Richard, R-Mercier, & Cellard, 2015). Most identified profiles in clinical psychology represent the quantitative-nomothetic tradition. The difficulty with using even normative based profiles to *predict* a diagnosis or other clinically meaningful outcomes is that there are a finite number of neurocognitive abilities overall, intelligence tests sample only a small set of these abilities, and there are many conditions, illnesses, disorders, and disabilities with overlapping profiles. As a result, the use of profiles, whether ipsatively created or based on general nomothetic patterns, is discouraged.

Despite a general discouragement of strongly interpreting observed profiles, intelligence tests remain one of our most studied and best predictors of clinically meaningful outcomes. For example, overall cognitive ability is typically the strongest predictor of school performance (Roth et al., 2015). When considering more targeted outcomes such as reading achievement, overall cognitive ability continues to be the strongest predictor. Adding neurocognitive interpretations to component or subtest scatter from an intelligence test tends to add little to predictive utility of general cognitive abilities (Watkins & Glutting, 2000). Adding information from other neurocognitive tests will often times add incremental utility to prediction. For example, adding measures of working memory from a different test improves the prediction of reading abilities (Krumm, Ziegler, & Buehner, 2008). Therefore, psychologists should focus on the level of a youth's cognitive ability and use other measures designed to assess specific neuropsychological deficits.

There are two reasons for the recommendation that neuropsychologists use specific neuropsychological assessments to evaluate specific neurocognitive abilities and not intelligence tests. First, neuropsychological assessments are often designed to evaluate deficits and are able to provide higher quality information regarding specific deficits. For example, the Delis–Kaplan Executive Functioning System (Delis, Kaplan, & Kramer, 2001) provides nine tests of executive functioning across presentation domains (i.e., verbal, visual) and types of executive functioning (e.g., fluency, inhibition, problem-solving, abstraction). Specific neuropsychological assessments allow for a more fine-grained analysis of neurocognitive abilities than intelligence tests. Second, intelligence tests have been selectively designed to measure broad cognitive abilities. Subtests on intelligence tests are saturated with variance from  $g$  (McGrew, LaForte, & Schrank, 2014; Wechsler, Raiford, & Holdnack, 2014) making it difficult to identify whether the remaining variance is reflective of error or specific abilities. Bifactor models tend to fit modern intelligence tests better than hierarchical models (Dombrowski, Canivez, & Watkins, 2017; Dombrowski, Canivez, Watkins, & Beaujean, 2015; Dombrowski, McGill, & Canivez, 2017b). Bifactor models consist of a general factor (i.e.,  $g$ ) that predicts each subtest as well as a set of orthogonal specific factors that predict domain specific subtests (e.g., working memory subtests are explained by both a general factor and a specific factor). More importantly, these analyses demonstrate that variance attributable to a

broad neurocognitive domain (e.g., working memory) is generally minimal and that most of the subtest variance can be attributed to *g*. Therefore, clinicians should use neurocognitive assessments designed to evaluate specific neurocognitive abilities because individual subtests tend to be saturated with *g* leaving minimal amounts of variance to be explained by a broad neurocognitive ability.

The most common type of standardized scores on intelligence tests are deviation-based IQ scores or Index scores. The standard score represents an individual child's relative standing compared to a large group of same-age peers. On most intelligence tests, scores are limited to a range of four standard deviations below or above the mean. Each intelligence test divides these scores into ranges and provides unique labels for these ranges. However, modern intelligence tests tend to use the same mean and standard deviation for standard scores. Therefore, some generalities about levels of intelligence can be broadly applied. Individuals with cognitive abilities two standard deviations below the mean (i.e., IQ < 70) tend to be identified early, have physical or genetic anomalies, have slower development across a broad range of development (e.g., communication skills, social skills), and will typically display significant academic, memory, attention, and rate of learning deficits relative to same-age peers (Bebko & Luhaorg, 1998; Feuerstein, Feuerstein, & Falik, 2010). Youth with cognitive abilities between one and two standard deviations below the mean (i.e., 70–85) tend to have difficulties with multistep or cognitively complex tasks, prefer concrete to abstract tasks, and are slower learners that require more practice than same-age peers (Masi, Marcheschi, & Pfanner, 1998). Youth with cognitive abilities one or more standard deviations above the mean (i.e., 115–130) tend to be academically inclined, more likely succeed in school, and more curious (Frisby, 2013). Therefore, neuropsychologists should make normative-based statements about a child's abilities based on intelligence test results.

Monitoring of long-term outcomes of individuals with different disorders, diseases, and disabilities is strongly recommended. For example, children with ADHD tend to grow out of ADHD with time. One hypothesis is that ADHD, its symptoms, and its neurocognitive deficits (e.g., executive dysfunction) could be due to delayed maturation of the brain. Longitudinal monitoring of neurocognitive abilities in youth with ADHD suggest that many delayed abilities tend to improve with time regardless of diagnostic status (Coghill, Hayward, Rhodes, Grimmer, & Matthews, 2014; Murray, Robinson, & Tripp, 2017). Similarly, neurocognitive abilities in healthy children also display heterogeneous trajectories, suggesting that uneven development of cognitive abilities is typical (van der Stel & Veenman, 2014; Walhovd et al., 2016). Longitudinal studies of educational interventions indicate improvement in specific abilities (i.e., raising poorer abilities) more than changing general cognitive abilities (Jordan, Wylie, & Mulhern, 2015; Ritchie, Bates, & Deary, 2015). However, making strong interpretations of change over time based on standard score placement is not recommended. As described in the quantitative-idiographic section, accounting for error in an individual's change score is preferred. Therefore, psychologists should not overinterpret differences in standard scores or relative placement to peers without determining whether the amount of change is reliable.

Intelligence tests are consistently among the most widely given measures by psychologists and are typically the top measure given by neuropsychologists (Rabin, Barr, & Burton, 2005). The ability of intelligence tests to *prescribe* a specific treatment is questionable and not widely studied. There are clear examples of where intelligence tests have value in the prescriptive process. For example, youth with low abilities demonstrate gains in cognitive abilities when prescribed stimulants (Marraccini, Weyandt, Rossi, & Gudmundsdottir, 2016). Youth with deficiencies in vitamin levels, iron, or iodine benefit from vitamin supplementation via a multivitamin, iron, or iodine supplement (Protzko, 2017). Additionally, lower cognitive abilities are consistently raised via learning to play a musical instrument (Protzko, 2017). In the context of psychotherapy, better cognitive abilities at the onset of treatment are associated with more improvement in psychotherapy than lower cognitive abilities (Mathiassen et al., 2012). Additionally, better cognitive abilities are associated with more improvement from more cognitively-oriented treatments (Bastos, Guimarães, & Trentini, 2017; Fournier et al., 2010), while lower cognitive abilities are associated with the use of behavioral interventions. Taken together, there is evidence that intelligence tests could be useful in helping to aid the *prescription* of treatment.

## **Review of two frequently used individually administered tests of intelligence for youth**

There are many different intelligence tests for children and adolescents that can be administered individually or in groups. Psychologists typically use individually administered intelligence tests for clinical assessments. For example, the Stanford–Binet-V (Roid, 2003), Differential Ability Scales-II (Elliot, 2007), and Reynolds Intellectual Assessment Scales (Reynolds & Kamphaus, 2003) are examples of individually administered intelligence tests for children and adolescents. The relevant information about their underlying theory, standardization, and interpretation may be found in their respective test manuals. However, two tests are typically among the most commonly used intelligence tests. The Wechsler Intelligence Scale for Children—fifth edition (WISC-V) and the Woodcock–Johnson Tests of Cognitive Abilities—fourth edition (WJ-IV). Additionally, these two tests have more substantial empirical literature bases evaluating their utility in clinical assessments. Therefore, brief reviews of the WISC-V and WJ-IV follow.

### **Wechsler Intelligence Scale for Children—fifth edition (WISC-V)**

#### ***Theory***

The WISC-V is an individually administered intelligence test for school-age children and adolescents. According to manual, the WISC-V was updated to (1)

increase the breadth of construct coverage (e.g., develop a new fluid reasoning subtest), (2) increase user friendliness (e.g., paper and electronic administration; reduce number of subtests required for FSIQ), (3) increase developmental appropriateness (e.g., reduce vocabulary level), (4) improve psychometric properties (e.g., update item and subtest scoring rules), and (5) enhance clinical utility (e.g., add subtests related to learning difficulties). However, the general decision to maintain much of Wechsler's tradition weakens the overall utility of the test. David Wechsler's working definition of intelligence for his tests was "intelligence is the overall capacity of an individual to understand and cope with the world around him" (Wechsler, 1974). During the original development of his tests, Wechsler did not first define the construct and then select tests that closely matched his definition. Instead, Wechsler defined his construct and combined verbal and performance tests from earlier intelligence tests into a single test. The early disconnect between theory and test continues to this day as modern editions of the WISC have updated the foundation Wechsler originally developed. For example, the WISC-V purports to measure five primary abilities instead of four on the WISC-IV or three on the WISC-R or two on the WISC. To measure these additional abilities, the WISC-V eliminated word reasoning and picture completion from the WISC-IV and added figure weights, visual puzzles and picture span as primary subtests. Additionally, the WISC-V added other subtests because they are sensitive to learning difficulties, not because they are part of a theory of intelligence. The WISC-V technical manual notes that the Wechsler series of tests are typically considered "atheoretical" by observers because no consistent underlying theory is put forth. The manual also argues that this is not accurate because many theoretical perspectives can be applied to the WISC-V *post hoc* (Wechsler et al., 2014). For example, subtests on the WISC-V can be arranged and analyzed using Cattell–Horn–Carroll (CHC) theory. The manual also points users towards alternative theories of intelligence as a framework for interpreting the WISC-V. Therefore, psychologists are encouraged to understand the empirical limits of the WISC-V and focus on the WISC-V's utility in aiding with *prediction, prescription, and process*.

## **Standardization**

The WISC-V was standardized on 2200 youth ranging in age from 6 years 0 months through 16 years 11 months. The youth were divided into 33 age bands, one band for each 4 month interval (e.g., 6 years, 0 months to 6 years, 3 months). Each age band consists of approximately 67 children. The WISC-V relied on stratified sampling to match the norming group to 2012 U.S. Census Bureau data with respect to age, gender, race/ethnicity, parent education level, and geographic region. Comparisons of the norming sample to the U.S. Census Data indicated that the WISC-V norming sample matched the overall U.S. population in terms of age, gender, race/ethnicity, parent education level, and geographic region very well. However, the norming sample under-sampled youth with specific learning disabilities, speech/language impairments, attention-deficit/hyperactivity disorder, and those who are gifted and talented. During development, youth with specific learning

disorders (i.e., reading, written expression, math), with ADHD, with disruptive behavior, who were English Language Learners, with autism with language impairment, with autism without language impairment, and with traumatic brain injuries were also sampled for specific comparisons. However, the norming sample excluded youth with sensory deficits, youth with a primary language other than English, and youth who displayed disruptive or noncompliant behavior. In summary, the WISC-V is well-matched to the general U.S. population for normal range assessment of intelligence, and caution must be used when using the WISC-V with youth with a primary language other than English (i.e., English Language Learners) or to youth with intellectual disabilities. For youth who are English Language Learners, testing in the youth's primary language is recommended. For youth with either extremely low or extremely high cognitive abilities, other tests are designed to more accurately assess abilities in the extreme ranges (e.g., Stanford-Binet V).

## ***Properties***

The WISC-V consists of 21 subtests. As seen in [Table 3.1](#), administration of all subtests could result in as many as 14 composite scales composed of overlapping subtests. However, the standard administration of the WISC-V consists of 10 scales and results in estimates on five 2 subtest composite indices and one overall estimate of Full Scale IQ. One of the major changes between the prior version and the current version is the number of scales that are needed to estimate the FSIQ. The WISC-V requires seven subtests (i.e., Block Design, Similarities, Matrix Reasoning, Digit Span, Coding, Vocabulary, and Figure Weights) to estimate the Full Scale IQ, an overall estimate of *g*. Three more subtests, for a total of 10 subtests (i.e., 7 for FSIQ plus Visual Puzzles, Picture Span, and Symbol Search) are required to estimate the five primary abilities indices. Therefore, the precise number of subtests required during a WISC-V administration is dependent upon the goal of the evaluation.

The WISC-V manual reports a hierarchical five factor structure. The Full Scale IQ, a measure of *g*, is the hierarchical factor. Specific factors are named indices. The five indices are: Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed. Note that the Perceptual Reasoning Index (PRI) from the WISC-IV no longer exists and was split into the Visual Spatial and Fluid Reasoning indices. However, the factor structure as presented in the manual is problematic. For example, the Arithmetic subtest does not appear to load strongly on any specific composite. Therefore, clinicians should be cautious in heavily interpreting the meaning of any index that includes Arithmetic.

The Full Scale IQ (FSIQ) is an overall measure a child's general cognitive abilities which include a person's ability to reason verbally, process visual-spatial information, use inductive and quantitative reasoning, temporarily store and use information (i.e., working memory), and process information quickly. The FSIQ is the most reliable ( $r_{xx} = .96$ ) of the composite scores and stable over a 1 month interval ( $r_{12} = .92$ ). In interpreting the WISC-V for most clinical purposes, preference should be given to the FSIQ over all other measures of ability because the FSIQ

has the strongest empirical base as it relates to Prediction, Prescription, and Process.

The Verbal Comprehension Index (VCI) is an overall measure of the ability to verbally reason and is heavily influenced by semantic knowledge. The VCI consists of Similarities, Vocabulary, Information, and Comprehension. Similarities measures verbal concept formation and abstract reasoning. Vocabulary measures word knowledge and verbal concept formation. Information measures one's ability to acquire, retain, and retrieve general factual knowledge. Comprehension is the most complex of these tests and measures because it requires verbal reasoning and conceptualization, verbal comprehension and expression, and practical knowledge and judgment. The VCI is reliable ( $r_{xx} = .92$ ) and stable over a one month interval ( $r_{12} = .94$ ).

The Visual Spatial Index (VSI) is an overall measure of visual spacing processing. The VSI consists of Block Design and Visual Puzzles. Block Design measures the ability to analyze and synthesize abstract visual stimuli with a motor component. Visual Puzzles measures the ability to mentally analyze and synthesize abstract visual stimuli without a motor component. The VSI is reliable ( $r_{xx} = .92$ ) although less stable over a one month interval relative to other composites ( $r_{12} = .84$ ).

The Fluid Reasoning Index (FRI) is an overall measure of inductive and quantitative reasoning. The FRI consists of Matrix Reasoning, Figure Weights, Picture Concepts, and Arithmetic. Matrix Reasoning measures the ability to identify underlying conceptual rules that link visual–spatial information. Figure Weights measures the ability to apply the quantitative concept of equality to visual images and apply concepts such as matching, addition, or multiplication. Picture Concepts measures the ability to identify underlying conceptual relationships among identifiable objects. Arithmetic is the most complicated of the FRI's subtests and measures a mixture of mental manipulation, concentration, attention, working memory, numerical reasoning ability, and short- and long-term memory. The FRI is the most reliable composite ( $r_{xx} = .93$ ) and least stable over a one month period ( $r_{12} = .75$ ).

The Working Memory Index (WMI) is an overall measure of working memory which is a cognitive system that temporarily holds and processes a limited amount of information. The WMI consists of Digit Span, Picture Span, and Letter–Number Sequencing. Digit Span measures auditory rehearsal, temporary storage capacity in working memory, and the ability to transform and manipulate numeric information. Picture Span measures visual working memory and visual working memory capacity. Letter–Number Sequencing measures auditory discrimination, auditory rehearsal, and the ability to transform and manipulate mental information. The WMI is reliable ( $r_{xx} = .92$ ) and stable over a one month period ( $r_{12} = .82$ ).

The Processing Speed Index (PSI) is an overall measure of the speed at which one processes information and makes simple decisions rapidly. The PSI consists of Coding, Symbol Search, and Cancellation. Coding measures processing speed, short-term visual memory, and psychomotor speed. Symbol Search measures visual–perceptual matching and decision-making speed. Cancellation measures the speed of visual–perceptual processing and decision-making.

### ***Useful details for additional interpretation***

Of note, the standard error of the measure can be manipulated in multiple manners. Each manipulation answers a different clinical question. [Table 3.2](#) presents clinically meaningful manipulations of this term for the WISC-V. First, one could ask, “How accurate is the obtained score of a person’s true score?” The WISC-V manual presents confidence intervals based on the  $SE_{measure}$ . These confidence intervals reflect the range in which an examinee’s true ability might fall. A clinician should interpret these values as: “Given Melinda’s performance on the FSIQ of 100, her true ability likely falls between 94 and 106 (95% confidence interval).” Second, a clinician might ask, “How consistent should this score be on repeated testing?” To estimate the accuracy of scores on repeated testing, one must apply the  $SE_{difference}$ . A clinician should interpret these values as: “On repeated testing, Melinda’s performance will likely fall between 92 and 108 (95% confidence interval).” Finally, a clinician might ask, “Has the person’s score changed significantly since last time they were tested?” The reliable change index could be used to determine whether scores on the measure have changed significantly over time ([Jacobson & Truax, 1991](#)). A clinician should interpret this comparison as: “Melinda’s earned an FSIQ of 85 today which is 15 points lower than her last FSIQ administration. Given that Melinda’s overall FSIQ changed by more than 8 points, her overall performance has significantly declined.”

### ***Critique***

The WISC-V continues the evolution of the WISC series and represents a substantial update over the previous version. However, independent analysis of the WISC-V has resulted in two substantial critiques of the test and information presented in the manuals. First, the WISC V’s factor structure is questionable. Changes in item content and subtests could result in changes in the expected factor structure ([Strauss, Spreen, & Hunter, 2000](#)). The WISC-V manual did not report tests for potential changes in the factor structure. Independent analyses of the WISC-V suggest that a four (collapsing VSI and FRI into Perceptual Reasoning) or five factor model fits reasonably well ([Canivez, Watkins, & Dombrowski, 2016](#); [Canivez, Watkins, & Dombrowski, 2017](#); [Dombrowski et al., 2015, 2017](#)). Second, the WISC-V manual presents traditional metrics of reliability that are biased when multidimensionality is present ([Raykov, 1997](#)). Independent analyses suggest that the FSIQ and PSI have adequate reliability for independent interpretation and that the VCI, VSI, FRI, and WMI do not contain enough reliable variance to be analyzed independently of FSIQ ([Canivez & Watkins, 2016](#)). Third, the preferred method for interpreting the WISC-V represents an *aptitude treatment interaction* (ATI) approach that focuses on personal strengths and weaknesses. As previously discussed, this approach in general is not recommended for four reasons: (1) the incremental utility of the comparisons is nearly nonexistent ([Kranzler & Floyd, 2013](#)), (2) the comparisons are not stable across time ([Borsuk, Watkins, & Canivez, 2006](#); [Watkins & Canivez, 2004](#)), (3) the comparisons are not stable across methods

**Table 3.2** Reliable change on the wechsler intelligence scales for children 5th edition

	Test–retest stability	$S_{diff}$	95% Reliable change
Verbal Comprehension Index	.94	5.20	11
Similarities	.88	1.47	3
Vocabulary	.90	1.34	3
Information	.88	1.47	3
Comprehension	.83	1.75	4
Visual Spatial Index	.84	8.49	17
Block Design	.81	1.85	4
Visual Puzzles	.80	1.90	4
Fluid Reasoning Index	.75	10.61	21
Matrix Reasoning	.78	1.99	4
Figure Weights	.82	1.80	4
Picture Concepts	.71	2.28	5
Arithmetic	.84	1.70	4
Working Memory Index	.82	9.00	18
Digit Span	.82	1.80	4
Picture Span	.80	1.90	4
Letter–Number Sequencing	.82	1.80	4
Processing Speed	.83	8.75	18
Coding	.81	1.85	4
Symbol Search	.80	1.90	4
Cancellation	.82	1.80	4
FSIQ	.92	6.00	12

for calculating them (Taylor, Miciak, Fletcher, & Francis, 2017), and (4) that ATI-based interventions are generally not effective unless paired with an achievement component (Fuchs et al., 2011). The critical reviews of the WISC-V are primarily focused on the interpretation of the WISC-V. According to its critics, clinicians should be careful not to overinterpret the results as the test likely only measures an overall general cognitive ability.

## Woodcock–Johnson Tests of Cognitive Abilities (WJ-IV COG)

### Theory

The WJ-IV COG is an individually administered intelligence test for individuals ages 2–90+ years. According to the manual, the WJ-IV COG was updated to (1) measure CHC theory more thoroughly (e.g., addition of new subtests and clusters), (2) revise and organize the overall battery (e.g., creation of the WJ-IV–Oral Language), (3) revise and update individual subtests to improve administration (e.g., simplify test administration procedures, additional items as needed

for more accurate ability estimates), and (4) update the normative data for the WJ-IV COG and conorm the WJ-IV COG, ACH, and OL for cross-battery comparisons. The WJ-IV COG and its related tests are explicitly based on modern interpretations of CHC theory (McGrew et al., 2014). The technical manual claims that the WJ-IV COG measures general intelligence (i.e.,  $g$ ) three different ways: (1) a seven test battery (General Intellectual Ability [GIA]), (2) three subtests (Brief Intellectual Ability [BIA]), and (3) four subtests ( $Gf-Gc$  Composite). The WJ-IV COG can also produce scores on the following broad CHC domains: Comprehension-Knowledge ( $Gc$ ), Fluid Reasoning ( $Gf$ ), Short-term Working Memory ( $Gwm$ ), Cognitive Processing Speed ( $Gs$ ), Auditory Processing ( $Ga$ ), Long-term Retrieval ( $Glr$ ), and Visual Processing ( $Gv$ ). Of note, the CHC-related factors are only measured via two subtests if the standard battery is administered. Therefore, the WJ-IV COG represents a test of intelligence based in modern CHC theory.

## **Standardization**

The WJ-IV COG was standardized on 7416 individuals between 2 years 0 months, and 90 + years. The school-age norming sample consisted of 3891 children in kindergarten through 12th grade. The youth were divided into age bands based on 12-month intervals (e.g., 12 years 0 months to 12 years 11 months). Each age band consisted of approximately 300 youth. The WJ-IV COG relied on stratified sampling to match the norming group to the 2010 U.S. Census Bureau data with respect to geographic region, sex, country of birth, race/ethnicity, community size, parent education level, and school type. Overall, the stratification approach resulted in a norming sample that matched the general U.S. population fairly well. Individual youth data was also weighted to account for minor discrepancies between sampling and the U.S. population (e.g., oversampled youth from the Midwest and West and youth with parents who had a high-school diploma; under-sampled youth from the South, youth with parents who had a college degree). The manual included a series of clinically important subsamples such as youth with learning disabilities (reading, math, and writing), traumatic brain injury, language delay, autism spectrum disorder, ADHD, gifted, and intellectual disability. Overall, the WJ-IV COG norming procedure produced a sample that is a close match to the general U.S. population in 2010.

The process of developing the WJ-IV COG is very well explained in the technical manual. Items and tests were initially tried out in smaller samples. Items were revised or new items were created as necessary to improve measurement across the range of cognitive ability. During development, the WJ-IV COG explicitly tested for three types of bias: male versus female, Hispanic versus nonHispanic, and white versus nonwhite. Overall, very few items displayed bias. In most cases, items displaying bias were removed. In some cases, biased items remained due to the item pool being too small. Overall, the WJ-IV COG made excellent progress in accounting for variance due to external factors (e.g., culture) on test performance.

## Properties

The WJ-IV COG consists of 18 subtests. As seen in [Table 3.3](#), administration of all subtests could result in as many as three general cognitive composite scales as well as additional seven CHC factors. However, the standard administration of the WJ-IV COG consists of 10 subtests and results in three estimates of general cognitive abilities (GIA, BIA, Gf–Gc Composite) and seven CHC factors. Please note that during the standard administration, most of the CHC factors are measured by a single subtest only. The manual explicitly states that the number of subtests administered should be reflective of the assessment question. Therefore, the precise number of subtests required during a WJ-IV COG administration is dependent upon the goal of the evaluation.

The WJ-IV manual reports a hierarchical factor structure. In the presented factor analyses, general cognitive ability (i.e.,  $g$ ) is measured via the GIA and a brief 3 subtest version (BIA) or a theoretically driven version (Gf–Gc) are also estimated. The seven broad CHC factors are hierarchically below the GIA estimate and they are: Gc, Gf, Gwm, Gs, Ga, Glr, and Gv. However, the factor structure as presented in the manual is problematic. The model supporting the structure outlined above was the best fitting model but the model had a poor fit to the norming data. Therefore, clinicians should be cautious in heavily interpreting the meaning of any given factor.

The GIA is an overall measure of a child's general cognitive abilities via sampling a single subtest from each of the seven CHC domains. The subtests are the first seven subtests administered, cognitively complex, and load strongly on both  $g$  and a CHC factor. The GIA represents a person's general ability to display acquired knowledge; reason with verbal and nonverbal information; perceive and manipulate information quickly; perceive, analyze, and synthesize visual and auditory patterns; and perform cognitive tasks rapidly. The GIA is the most reliable ( $r_{xx} = .96$ ) of the composite scores. In interpreting the WJ-IV COG, preference should be given to the GIA over all other measures of ability.

Interpretation of the CHC factors on the WJ-IV COG is more complicated than on other intelligence tests. If only the standard battery is administered, then the CHC factors for Gs, Ga, Glr, and Gv represent the same score as a single subtest. If the extended battery is administered, then the CHC factors represent a combined score of two subtests. Therefore, during standard battery administrations, clinicians should be wary of strongly interpreting the CHC factors.

The Gc factor is an overall measure of a child's knowledge and skills valued by U.S. culture. The Gc factor consists of Oral Vocabulary and General Information. Oral Vocabulary measures verbal concept formation and abstract reasoning by asking for synonyms and antonyms. General Information measures one's ability to acquire, retain, and retrieve general factual knowledge.

The Gf factor is an overall measure of a child's ability to deliberately and flexibly control attention to solve novel problems. The Gf factor consists of Number Series, Concept Formation, and Analysis–Synthesis. Number Series measures the ability to represent and manipulate points on a mental number line and identify and

**Table 3.3** Woodcock–johnson IV test of cognitive abilities subscales and indices

		General intellectual ability	Brief intellectual ability	Gf–Gc composite	Comprehension-knowledge	Fluid reasoning	Short-term working memory	Cognitive processing speed	Auditory processing	Long-term retrieval	Visual processing
		GIA	BIA		Gc	Gf	Gwm	Gs	Ga	Glr	Gv
1	Oral Vocabulary	P	P	P	P	P					
2	Number Series	P	P	P							
3	Verbal Attention	P	P								
4	Letter-Pattern Matching	P									
5	Phonological Processing	P									
6	Story Recall	P									
7	Visualization	P									
8	General Information			P	P						
9	Concept Formation			P		P					
10	Numbers Reversed					P	P				

11	Number-Pattern Matching							P		
12	Nonword Repetition									
13	Visual–Auditory Learning								P	
14	Picture Recognition									P
15	Analysis–Synthesis				S		S			
16	Object–Number Sequencing									
17	Pair Cancellation						P			
18	Memory for Words <sup>a</sup>									

<sup>a</sup>Loads only to a narrow ability cluster, does not load to a cognitive composite or CHC factor.

apply a rule. Concept Formation measures the ability to apply rule-based categorization including rule-switching and making inferences. Analysis–Synthesis measures the ability to reason and deduce patterns.

The *Gwm* factor is an overall measure of a child's ability to encode, maintain, and manipulate information in immediate awareness. The *Gwm* factor consists of Verbal Attention, Numbers Reversed, and Object–Number Sequencing. Verbal Attention measures working memory capacity by selectively recalling meaningful verbal information interspersed with distracting information. Numbers Reversed measures auditory rehearsal and temporary storage capacity in working memory. Object–Number Sequencing measures auditory rehearsal and the ability to transform and manipulate mental information.

The *Gs* factor is an overall measure of a child's ability to perform simple and complex cognitive tasks rapidly. The *Gs* factor consists of Letter–Pattern Matching and Pair Cancellation. Letter–Pattern Matching measures speeded visual matching. Pair Cancellation measures the ability to rapidly control attention and sustain attention while finding similar visual information.

The *Ga* factor is an overall measure of a child's ability to discriminate meaningful nonverbal sound information. The *Ga* factor consists of Phonological Processing and Nonword Repetition. Phonological Processing measures both overall oral word knowledge and speed of lexical access. Nonword Repetition measures the ability to sequence phonological elements orally.

The *Glr* factor is an overall measure of a child's ability to store, consolidate, and retrieve information over time. *Glr* is different from *Gwm*. *Glr* requires displacement of a memory from primary memory to storage whereas as *Gwm* maintains a memory in primary memory. The *Glr* factor consists of Story Recall and Visual–Auditory Learning. Story Recall measures the ability to remember contextualized verbal information over time. Visual–Auditory Learning measures the ability to encode and store novel information over time.

The *Gv* factor is an overall measure of a child's ability to solve problems with mental imagery. The *Gv* factor consists of Visualization and Picture Recognition. Visualization measures the ability to identify visual features and mentally rotate images. Picture Recognition measures the ability to match visual stimuli with previously learned visual stimuli.

### *Critique*

The WJ-IV COG continues the evolution of the WJ series and represents a meaningful update over the previous version as well as a theoretical statement regarding the nature of intelligence. However, independent analyses of the WJ-IV COG has resulted in substantial critiques of the structure of the test and the recommended method for interpreting test results. First, the WJ-IV COG's factor structure is questionable. The manual presents a seven factor model informed by CHC theory and this model is the best fitting model of the limited set of models presented in the manual. However, the CHC informed structure represents poor fit of the standardization sample data indicating that CHC theory as operationalized in

the WJ-IV COG is not measured adequately. Independent analyses of the WJ-IV COG suggest that fewer factors in addition to  $g$  are measured on the WJ-IV COG (Dombrowski, McGill, & Canivez, 2017a; Dombrowski et al., 2017b). For example, one analysis indicates that the WJ-IV COG might best represent measures of: (1) visual processing ( $Gv$ ), (2) processing speed ( $Gs$ ), (3) working memory ( $Gwm$ ), and (4) crystallized intelligence ( $Gc$ ), with fluid reasoning ( $Gf$ ) subtests primarily loading on the overall measure of intelligence ( $g$ ) and not separating to a broad CHC factor (Dombrowski et al., 2017a). Second, the WJ-IV COG manual presents traditional metrics of reliability that are biased when multidimensionality due to higher order factors is present (Raykov, 1997). Independent analyses of the WJ-IV COG suggest that only GIA has adequate reliability for independent interpretation after accounting for the seven CHC factors and that the seven CHC factors do not contain enough reliable variance to be analyzed independently (Dombrowski et al., 2017a). Third, the preferred method for interpreting the WJ-IV COG represents an *aptitude treatment interaction* approach that focuses on personal strengths and weaknesses. As previously discussed, this approach is generally not recommended (Taylor et al., 2017). In the technical manual, a series of comparisons between specifically identified clinical groups highlights the weakness of within person comparisons. The manual only presents selective subtests so it is difficult to discuss overall profiles. However, of the presented profiles many clinical groups displayed similar profiles on one or more of the subtests (e.g., youth with learning disabilities [reading, math, and writing], traumatic brain injury, language delay, and autism spectrum disorder all had mean scores of approximately 80 on Oral Vocabulary). In summary, the critical reviews of the WJ-IV COG raise meaningful concerns about the overall structure of the test that directly influence how it is interpreted. Therefore, clinicians should be careful not to overinterpret the results as the test likely only measures an overall general cognitive ability and likely does not reliably measure other aspects of cognition.

## References

- Agrell, B., & Dehlin, O. (2012). The clock-drawing test. 1998. *Age and Ageing*, 41(Suppl 3), iii41–45. Available from <https://doi.org/10.1093/ageing/afs149>.
- Arrington, C. N., Ware, A. L., Ahmed, Y., Kulesz, P. A., Dennis, M., & Fletcher, J. M. (2016). Are shunt revisions associated with IQ in congenital hydrocephalus? A meta-analysis. *Neuropsychology Review*, 26(4), 329–339. Available from <https://doi.org/10.1007/s11065-016-9335-z>.
- Bastos, A. G., Guimarães, L. S., & Trentini, C. M. (2017). Predictors of response in the treatment of moderate depression. *Revista Brasileira de Psiquiatria*, 39(1), 12–20. Available from <https://doi.org/10.1590/1516-4446-2016-1976>.
- Beaujean, A. A. (2015). John Carroll's views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence*, 3(4), 121–136. Available from <https://doi.org/10.3390/jintelligence3040121>.

- Bebko, J. M., & Luhorg, H. (1998). The development of strategy use and metacognitive processing in mental retardation: Some sources of difficulty. In J. A. Burack, R. M. Hodapp, E. Zigler, J. A. Burack, R. M. Hodapp, & E. Zigler (Eds.), *Handbook of mental retardation and development* (pp. 382–407). New York, NY, US: Cambridge University Press.
- Binet, A., & Simon, T. (2003). *New methods for the diagnosis of the intellectual level of subnormals. The history of psychology: Fundamental questions* (pp. 270–287). New York, NY, US: Oxford University Press.
- Boake, C. (2002). From the Binet–Simon to the Wechsler–Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, 24(3), 383–405. Available from <https://doi.org/10.1076/jcen.24.3.383.981>.
- Borsuk, E. R., Watkins, M. W., & Canivez, G. L. (2006). Long-term stability of membership in a Wechsler intelligence scale for children-third edition (WISC-III) subtest core profile taxonomy. *Journal of Psychoeducational Assessment*, 24(1), 52–68. Available from <https://doi.org/10.1177/0734282905285225>.
- Bossuyt, P. M. M., Reitsma, J. B., Linnet, K., & Moons, K. G. M. (2012). Beyond diagnostic accuracy: The clinical utility of diagnostic tests. *Clinical Chemistry*, 58(12), 1636–1643. Available from <https://doi.org/10.1373/clinchem.2012.182576>.
- Brydges, C. R., Landes, J. K., Reid, C. L., Campbell, C., French, N., & Anderson, M. (2018). Cognitive outcomes in children and adolescents born very preterm: A meta-analysis. *Developmental Medicine & Child Neurology*. Available from <https://doi.org/10.1111/dmcn.13685>.
- Burns, M. K., Petersen-Brown, S., Haegele, K., Rodriguez, M., Schmitt, B., Cooper, M., & VanDerHeyden, A. M. (2016). Meta-analysis of academic interventions derived from neuropsychological data. *School Psychology Quarterly*, 31(1), 28–42. Available from <https://doi.org/10.1037/spq0000117>.
- Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test–retest correlations. *The Clinical Neuropsychologist*, 27 (7), 1077–1105. Available from <https://doi.org/10.1080/13854046.2013.809795>.
- Canivez, G. L., & Watkins, M. W. (2016). *Review of the Wechsler intelligence scale for children-fifth edition: Critique, commentary, and independent analyses. Intelligent Testing with the WISC-V* (1st ed., pp. 683–702). Hoboken, NJ: Wiley.
- Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the Wechsler intelligence scale for children-fifth edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment*, 28(8), 975–986. Available from <https://doi.org/10.1037/pas0000238>.
- Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler intelligence scale for children-fifth edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment*, 29(4), 458–472. Available from <https://doi.org/10.1037/pas0000358>.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.
- Coghill, D. R., Hayward, D., Rhodes, S. M., Grimmer, C., & Matthews, K. (2014). A longitudinal examination of neuropsychological and clinical functioning in boys with attention deficit hyperactivity disorder (ADHD): Improvements in executive functioning do not explain clinical improvement. *Psychological Medicine*, 44(5), 1087–1099. Available from <https://doi.org/10.1017/S0033291713001761>.
- Daleiden, E., Drabman, R. S., & Benton, J. (2002). The guide to the assessment of test session behavior: Validity in relation to cognitive testing and parent-reported behavior

- problems in a clinical sample. *Journal of Clinical Child and Adolescent Psychology*, 31(2), 263–271. Available from <https://doi.org/10.1207/153744202753604539>.
- Das, J. P., Cummins, J., Kirby, J. R., & Jarman, R. F. (1979). Simultaneous and successive processes, language and mental abilities. *Canadian Psychological Review/Psychologie Canadienne*, 20(1), 1–11. Available from <https://doi.org/10.1037/h0081488>.
- Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system (D-KEFS)*. San Antonio, TX: The Psychological Corporation.
- De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, 141(4), 858–900. Available from <https://doi.org/10.1037/a0038498>.
- De Ruiter, M. A., Van Mourik, R., Schouten-Van Meeteren, A. Y. N., Grootenhuis, M. A., & Oosterlaan, J. (2013). Neurocognitive consequences of a paediatric brain tumour and its treatment: A meta-analysis. *Developmental Medicine & Child Neurology*, 55(5), 408–417. Available from <https://doi.org/10.1111/dmcn.12020>.
- Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2017). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology*. Available from <https://doi.org/10.1007/s40688-017-0125-2>.
- Dombrowski, S. C., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2015). Exploratory bifactor analysis of the Wechsler intelligence scale for children-fifth edition with the 16 primary and secondary subtests. *Intelligence*, 53, 194–201. Available from <https://doi.org/10.1016/j.intell.2015.10.009>.
- Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2017a). Exploratory and hierarchical factor analysis of the WJ-IV cognitive at school age. *Psychological Assessment*, 29(4), 394–407. Available from <https://doi.org/10.1037/pas0000350>.
- Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2017b). Hierarchical exploratory factor analyses of the Woodcock–Johnson IV full test battery: Implications for CHC application in school psychology. *School Psychology Quarterly*. Available from <https://doi.org/10.1037/spq0000221>.
- Donders, J. (Ed.), (2016). *Neuropsychological report writing*. New York: Guilford. Retrieved from: <https://www.guilford.com/books/Neuropsychological-Report-Writing/Jacobus-Donders/9781462524174>.
- Ek, U., Fernell, E., Westerlund, J., Holmberg, K., Olsson, P.-O., & Gillberg, C. (2007). Cognitive strengths and deficits in schoolchildren with ADHD. *Acta Paediatrica (Oslo, Norway: 1992)*, 96(5), 756–761. Available from <https://doi.org/10.1111/j.1651-2227.2007.00297.x>.
- Ekman, P. (1984). Expression and the nature of emotion. In K. Scherer, & P. Ekman (Eds.), *Approaches to emotion*. Hillsdale, NJ: Lawrence Erlbaum.
- Elliot, C. D. (2007). *Differential ability scales* (2nd ed.). San Antonio, TX: Harcourt Assessment.
- Esquirol, J. É. D. (2012). *Mental maladies: Treatise on insanity* (Vol. 1). Forgotten Books.
- Feifer, S. G. (2008). Integrating Response to Intervention (RTI) with neuropsychology: A scientific approach to reading. *Psychology in the Schools*, 45(9), 812–825. Available from <https://doi.org/10.1002/pits.20328>.
- Feuerstein, R., Feuerstein, R. S., & Falik, L. H. (2010). *Beyond smarter: Mediated learning and the brain's capacity for change*. New York, NY, US: Teachers College Press.
- Fiorello, C. A., Hale, J. B., & Snyder, L. E. (2006). Cognitive hypothesis testing and response to intervention for children with reading problems. *Psychology in the Schools*, 43(8), 835–853. Available from <https://doi.org/10.1002/pits.20192>.

- Fiorello, C. A., Hale, J. B., & Wycoff, K. L. (2012). *Cognitive hypothesis testing: Linking test results to the real world. Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.). Guilford.
- Flanagan, D. P., & Alfonso, V. (2017). *Essentials of WISC-V assessment*. Hoboken, NJ: Wiley Retrieved from: . Available from <https://www.wiley.com/en-us/Essentials+of+WISC+V+Assessment-p-9781118980873>.
- Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity: A patient-level meta-analysis. *JAMA*, 303(1), 47–53. Available from <https://doi.org/10.1001/jama.2009.1943>.
- Frazier, T. W., Demaree, H. A., & Youngstrom, E. A. (2004). Meta-analysis of intellectual and neuropsychological test performance in attention-deficit/hyperactivity disorder. *Neuropsychology*, 18(3), 543–555. Available from <https://doi.org/10.1037/0894-4105.18.3.543>.
- Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*, 35(2), 169–182. Available from <https://doi.org/10.1016/j.intell.2006.07.002>.
- Fridlund, A. J. (1991). Evolution and facial action in reflex, social motive, and paralanguage. *Biological Psychology*, 32(1), 3–100. Available from [https://doi.org/10.1016/0301-0511\(91\)90003-Y](https://doi.org/10.1016/0301-0511(91)90003-Y).
- Frisby, C. (2013). *General cognitive ability, learning and instruction. Meeting the psychoeducational needs of minority children: Evidence-based guidelines for school psychologists and other school personnel*. New York, NY: Wiley.
- Fuchs, D., Hale, J. B., & Kearns, D. M. (2011). On the importance of a cognitive processing perspective: An introduction. *Journal of Learning Disabilities*, 44(2), 99–104. Available from <https://doi.org/10.1177/0022219411400019>.
- Glutting, J. J., McDermott, P. A., Konold, T. R., Snelbaker, A. J., & Watkins, M. W. (1998). More ups and downs of subtest analysis: Criterion validity of the DAS with an unselected cohort. *School Psychology Review*, 27(4), 599–612.
- Glutting, J. J., McDermott, P. A., Watkins, M. M., & Kush, J. C. (1997). The base rate problem and its consequences for interpreting children's ability profiles. *School Psychology Review*, 26(2), 176–188.
- Glutting, J. J., Oakland, T., & Konold, T. R. (1994). Criterion-related bias with the guide to the assessment of test-session behavior for the WISC-III and WIAT: Possible race/ethnicity, gender, and SES effects. *Journal of School Psychology*, 32(4), 355–369. Available from [https://doi.org/10.1016/0022-4405\(94\)90033-7](https://doi.org/10.1016/0022-4405(94)90033-7).
- Glutting, J. J., Youngstrom, E. A., Oakland, T., & Watkins, M. W. (1996). Situational specificity and generality of test behaviors for samples of normal and referred children. *School Psychology Review*, 25(1), 94–107.
- Hale, J. B., Fiorello, C. A., Bertin, M., & Sherman, R. (2003). Predicting math achievement through neuropsychological interpretation of WISC-III variance components. *Journal of Psychoeducational Assessment*, 21(4), 358–380. Available from <https://doi.org/10.1177/073428290302100404>.
- Hebben, N., & Milberg, W. (2009). *Essentials of neuropsychological assessment* (2nd ed.). Hoboken, NJ, US: John Wiley & Sons Inc.
- Hinton-Bayre, A. D. (2016). Clarifying discrepancies in responsiveness between reliable change indices. *Archives of Clinical Neuropsychology*, 31(7), 754–768. Available from <https://doi.org/10.1093/arclin/acw064>.

- Humphreys, L. G. (1994). Intelligence from the standpoint of a (pragmatic) behaviorist. *Psychological Inquiry*, 5(3), 179–192. Available from [https://doi.org/10.1207/s15327965pli0503\\_1](https://doi.org/10.1207/s15327965pli0503_1).
- 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(1), 12–19.
- Jordan, J. A., Wylie, J., & Mulhern, G. (2015). Individual differences in children's paths to arithmetical development. In R. C. Kadosh, A. Dowker, R. C. Kadosh, & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 975–992). New York, NY, US: Oxford University Press. Available from <https://doi.org/10.1093/oxfordhb/9780199642342.013.015>.
- Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2012). A history of intelligence test interpretation. In D. P. Flanagan, P. L. Harrison, D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 56–70). New York, NY, US: Guilford Press.
- Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll, & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function: Research, measurement, and practice* (pp. 125–167). Washington, DC: American Psychological Association.
- Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). New York: Wiley.
- Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent testing with the WISC-V*. Hoboken, NJ: Wiley. Retrieved from: [https://www.amazon.com/Intelligent-Testing-WISC-V-Alan-Kaufman/dp/1118589238/ref=sr\\_1\\_1?ie=UTF8&qid=1506897828&sr=8-1&keywords=intelligent+testing+with+the+wisc-v](https://www.amazon.com/Intelligent-Testing-WISC-V-Alan-Kaufman/dp/1118589238/ref=sr_1_1?ie=UTF8&qid=1506897828&sr=8-1&keywords=intelligent+testing+with+the+wisc-v).
- Kazdin, A. E. (2001). Almost clinically significant ( $p < .10$ ): Current measures may only approach clinical significance. *Clinical Psychology: Science and Practice*, 8(4), 455–462. Available from <https://doi.org/10.1093/clipsy/8.4.455>.
- Königs, M., Engenhorst, P. J., & Oosterlaan, J. (2016). Intelligence after traumatic brain injury: Meta-analysis of outcomes and prognosis. *European Journal of Neurology*, 23 (1), 21–29. Available from <https://doi.org/10.1111/ene.12719>.
- Kranzler, J. H., & Floyd, R. G. (2013). *Assessing intelligence in children and adolescents: A practical guide (Layflat edition)*. New York: The Guilford Press.
- Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, B., & Thibodaux, L. (2016). Classification agreement analysis of cross-battery assessment in the identification of specific learning disorders in children and youth. *International Journal of School & Educational Psychology*, 4(3), 124–136. Available from <https://doi.org/10.1080/21683603.2016.1155515>.
- Krumm, S., Ziegler, M., & Buehner, M. (2008). Reasoning and working memory as predictors of school grades. *Learning and Individual Differences*, 18(2), 248–257. Available from <https://doi.org/10.1016/j.lindif.2007.08.002>.
- Legg, S., & Hutter, M. (2007). *A collection of definitions of intelligence. Advances in artificial general intelligence: Concepts, architectures and algorithms: Proceedings of the AGI workshop 2006*. Amsterdam, Netherlands: IOS Press.
- Lezak, M. D. (1988). IQ: R.I.P. *Journal of Clinical and Experimental Neuropsychology*, 10 (3), 351–361. Available from <https://doi.org/10.1080/01688638808400871>.
- Livingston, R. B., Jennings, E., Reynolds, C. R., & Gray, R. M. (2003). Multivariate analyses of the profile stability of intelligence tests: High for IQs, low to very low for subtest analyses. *Archives of Clinical Neuropsychology*, 18(5), 487–507. Available from [https://doi.org/10.1016/S0887-6177\(02\)00147-6](https://doi.org/10.1016/S0887-6177(02)00147-6).

- Maassen, G. H. (2004). The standard error in the Jacobson and Truax reliable change index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society: JINS*, 10(6), 888–893.
- Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "intelligent testing" approach to the WISC-III. *School Psychology Quarterly*, 12(3), 197–234. Available from <https://doi.org/10.1037/h0088959>.
- Malarbi, S., Abu-Rayya, H. M., Muscara, F., & Stargatt, R. (2017). Neuropsychological functioning of childhood trauma and post-traumatic stress disorder: A meta-analysis. *Neuroscience and Biobehavioral Reviews*, 72, 68–86. Available from <https://doi.org/10.1016/j.neubiorev.2016.11.004>.
- Marraccini, M. E., Weyandt, L. L., Rossi, J. S., & Gudmundsdottir, B. G. (2016). Neurocognitive enhancement or impairment? A systematic meta-analysis of prescription stimulant effects on processing speed, decision-making, planning, and cognitive perseveration. *Experimental and Clinical Psychopharmacology*, 24(4), 269–284. Available from <https://doi.org/10.1037/pha0000079>.
- Masi, G., Marcheschi, M., & Pfanner, P. (1998). Adolescents with borderline intellectual functioning: Psychopathological risk. *Adolescence*, 33(130), 415–424.
- Masson, M., Bussières, E.-L., East-Richard, C., R-Mercier, A., & Cellard, C. (2015). Neuropsychological profile of children, adolescents and adults experiencing maltreatment: A meta-analysis. *The Clinical Neuropsychologist*, 29(5), 573–594. Available from <https://doi.org/10.1080/13854046.2015.1061057>.
- Mathiassen, B., Brøndbo, P. H., Waterloo, K., Martinussen, M., Eriksen, M., Hanssen-Bauer, K., & Kvernmo, S. (2012). IQ as a moderator of outcome in severity of children's mental health status after treatment in outpatient clinics. *Child and Adolescent Psychiatry and Mental Health*, 6. Available from <https://doi.org/10.1186/1753-2000-6-22>.
- McCloskey, G., & Maerlender, A. (2005). The WISC-IV integrated. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical use and interpretation: Scientist-practitioner perspectives* (pp. 101–149). Elsevier. Retrieved from: <https://doi.org/10.1016/B978-012564931-5/50005-0>.
- McConaughy, S. H. (2005). Direct observational assessment during test sessions and child clinical interviews. *School Psychology Review*, 34(4), 490–506.
- McConaughy, S. H., & Achenbach, T. M. (2004). *Manual for the test observation form*. Burlington, VT: University of Vermont, Center for Children, Youth, & Families.
- McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment*, 8 (3), 290–302. Available from <https://doi.org/10.1177/07342899000800307>.
- McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *The Journal of Special Education*, 25(4), 504–526. Available from <https://doi.org/10.1177/002246699202500407>.
- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37(1), 1–10. Available from <https://doi.org/10.1016/j.intell.2008.08.004>.
- McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical manual*. Rolling Meadows, IL: Riverside.
- Meehl, P. E. (1997). Credentialled persons, credentialled knowledge. *Clinical Psychology: Science and Practice*, 4(2), 91–98. Available from <https://doi.org/10.1111/j.1468-2850.1997.tb00103.x>.

- Murray, A. L., Robinson, T., & Tripp, G. (2017). Neurocognitive and symptom trajectories of ADHD from childhood to early adolescence. *Journal of Developmental and Behavioral Pediatrics*, 38(7), 465–475. Available from <https://doi.org/10.1097/DBP.0000000000000476>.
- Naglieri, J. A. (2000). Intelligence testing in the 21st century: A look at the past and suggestion for the future. *Educational and Child Psychology*, 17(3), 6–18.
- Neisser, U., Boodoo, G., Bouchard, T. J. J., Boykin, A. W., Brody, N., Ceci, S. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51(2), 77–101. Available from <https://doi.org/10.1037/0003-066X.51.2.77>.
- Oakland, T., Callueng, C., & Harris, J. G. (2012). The impact of test-taking behaviors on WISC-IV Spanish domain scores in its standardization sample. *Journal of Psychoeducational Assessment*, 30(2), 139–147. Available from <https://doi.org/10.1177/0734282911423358>.
- Oh, H.-J., Glutting, J. J., & McDermott, P. A. (1999). An epidemiological-cohort study of DAS processing speed factor: How well does it identify concurrent achievement and behavior problems? *Journal of Psychoeducational Assessment*, 17(4), 362–375. Available from <https://doi.org/10.1177/07342829901700406>.
- Pievsky, M. A., & McGrath, R. E. (2018). The neurocognitive profile of attention-deficit/hyperactivity disorder: A review of meta-analyses. *Archives of Clinical Neuropsychology*, 33(2), 143–157. Available from <https://doi.org/10.1093/arclin/acx055>.
- Poreh, A. M. (2006). Methodological quandaries of the quantified process approach. In A. M. Poreh (Ed.), *The quantified process approach to neuropsychological assessment* (pp. 27–41). New York, NY: Taylor & Francis.
- Postal, K., Chow, C., Jung, S., Erickson-Moreo, K., Geier, F., & Lanca, M. (2018). The stakeholders' project in neuropsychological report writing: a survey of neuropsychologists' and referral sources' views of neuropsychological reports. *The Clinical Neuropsychologist*, 32(3), 326–344. Available from <https://doi.org/10.1080/13854046.2017.1373859>.
- Protzko, J. (2017). Raising IQ among school-aged children: Five meta-analyses and a review of randomized controlled trials. *Developmental Review*, 46, 81–101. Available from <https://doi.org/10.1016/j.dr.2017.05.001>.
- 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(1), 33–65. Available from <https://doi.org/10.1016/j.acn.2004.02.005>.
- Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32(4), 329–353. Available from [https://doi.org/10.1207/s15327906mbr3204\\_2](https://doi.org/10.1207/s15327906mbr3204_2).
- Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds intellectual assessment scales*. Lutz, FL: Psychological Assessment Resources.
- Ritchie, S. J., Bates, T. C., & Deary, I. J. (2015). Is education associated with improvements in general cognitive ability, or in specific skills? *Developmental Psychology*, 51(5), 573–582. Available from <https://doi.org/10.1037/a0038981>.
- Robinson, T. M. (1972). *Plato's psychology*. Toronto: University of Toronto Press.
- Roid, G. H. (2003). *Stanford-Binet intelligence scales, fifth edition, Technical manual* (5th ed.). Itasca, IL: Riverside.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118–137. Available from <https://doi.org/10.1016/j.intell.2015.09.002>.
- Sala, G., & Gobet, F. (2017). Working memory training in typically developing children: A meta-analysis of the available evidence. *Developmental Psychology*, 53(4), 671–685. Available from <https://doi.org/10.1037/dev0000265>.

- Sattler, J. M. (2008). *Assessment of children: Cognitive foundations, 5th Edition* (5th edition). San Diego: Jerome M. Sattler, Publisher.
- Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed, pp. 99–114). New York, N.Y: Guilford.
- Semrud-Clikeman, M., Wilkinson, A., & Wellington, T. M. (2005). Evaluating and using qualitative approaches to neuropsychological assessment. In R. C. D'Amato, E. Fletcher-Janzen, C. R. Reynolds, R. C. D'Amato, E. Fletcher-Janzen, & C. R. Reynolds (Eds.), *Handbook of school neuropsychology* (pp. 287–302). Hoboken, NJ, US: John Wiley & Sons Inc.
- Spearman, C. (1904). “General Intelligence,” objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–292. Available from <https://doi.org/10.2307/1412107>.
- Sternberg, R. J. (1990). Metaphors of mind: Conceptions of the nature of intelligence (*First Thus edition*). Cambridge: Cambridge University Press.
- Straus, S., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2010). *Evidence-based medicine: how to practice and teach it* (4th ed). Churchill Livingstone Retrieved from . Available from <https://www.elsevier.com/books/evidence-based-medicine/straus/978-0-7020-3127-4>.
- Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment*, 12(3), 237–244. Available from <https://doi.org/10.1037/1040-3590.12.3.237>.
- Stuebing, K. K., Barth, A. E., Trahan, L. H., Reddy, R. R., Miciak, J., & Fletcher, J. M. (2015). Are child cognitive characteristics strong predictors of responses to intervention? A meta-analysis. *Review of Educational Research*, 85(3), 395–429. Available from <https://doi.org/10.3102/0034654314555996>.
- Stuebing, K. K., Fletcher, J. M., LeDoux, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal*, 39(2), 469–518. Available from <https://doi.org/10.3102/00028312039002469>.
- Taylor, W. P., Miciak, J., Fletcher, J. M., & Francis, D. J. (2017). Cognitive discrepancy models for specific learning disabilities identification: Simulations of psychometric limitations. *Psychological Assessment*, 29(4), 446–457. Available from <https://doi.org/10.1037/pas0000356>.
- Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, 5(4), 357–369.
- Thorndike, R., & Lohman, D. F. (1989). *Century of ability testing*. Chicago: Riverside Pub Co.
- Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, 41(1), 1–32. Available from <https://doi.org/10.1037/h0075959>.
- Tigner, R. B., & Tigner, S. S. (2000). Triarchic theories of intelligence: Aristotle and Sternberg. *History of Psychology*, 3(2), 168–176. Available from <https://doi.org/10.1037/1093-4510.3.2.168>.
- Tzotzoli, P. (2012). A guide to neuropsychological report writing. *Health*, 04(10), 821. Available from <https://doi.org/10.4236/health.2012.410126>.
- U.S. Department of Education. (2016). Digest of education statistics, 2015 NCES 2016-014 (No. 2015). Washington DC: U.S. Department of Education. Retrieved from [https://nces.ed.gov/programs/digest/d16/tables/dt16\\_204.30.asp?current=yes](https://nces.ed.gov/programs/digest/d16/tables/dt16_204.30.asp?current=yes).

- van der Maas, H. L. J., Kan, K.-J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. seriously. *Journal of Intelligence*, 2(1), 12–15. Available from <https://doi.org/10.3390/jintelligence2010012>.
- van der Stel, M., & Veenman, M. V. J. (2014). Metacognitive skills and intellectual ability of young adolescents: A longitudinal study from a developmental perspective. *European Journal of Psychology of Education*, 29(1), 117–137. Available from <https://doi.org/10.1007/s10212-013-0190-5>.
- Walhovd, K. B., Krogsrud, S. K., Amlien, I. K., Bartsch, H., Bjørnerud, A., Due-Tønnessen, P., ... Fjell, A. M. (2016). Neurodevelopmental origins of lifespan changes in brain and cognition. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 113(33), 9357–9362.
- Watkins, M. W. (2003). IQ subtest analysis: clinical acumen or clinical illusion? *The Scientific Review of Mental Health Practice: Objective Investigations of Controversial and Unorthodox Claims in Clinical Psychology, Psychiatry, and Social Work*, 2(2), 118–141.
- Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of wisc-iii subtest composite: strengths and weaknesses. *Psychological Assessment*, 16(2), 133–138. Available from <https://doi.org/10.1037/1040-3590.16.2.133>.
- Watkins, M. W., & Glutting, J. J. (2000). Incremental validity of WISC-III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment*, 12(4), 402–408. Available from <https://doi.org/10.1037/1040-3590.12.4.402>.
- Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore: The Williams & Wilkins Company.
- Wechsler, D. (1974). *Wechsler intelligence scale for children - revised*. New York, NY: Psychological Corporation.
- Wechsler, D., & Kaplan, E. (2015). *WISC-V integrated: Technical and interpretive manual*. Bloomington: PsychCorp.
- Wechsler, D., Raiford, S. E., & Holdnack, J. A. (2014). *WISC-V: Technical and interpretative manual*. Bloomington: Pearson ClinicalAssessment.
- Weicker, J., Villringer, A., & Thöne-Otto, A. (2016). Can impaired working memory functioning be improved by training? A meta-analysis with a special focus on brain injured patients. *Neuropsychology*, 30(2), 190–212. Available from <https://doi.org/10.1037/neu0000227>.
- Werner, H. (1937). Process and achievement: A basic problem of education and developmental psychology. *Harvard Educational Review*, 7, 353–368.
- Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment*, 82(1), 50–59. Available from [https://doi.org/10.1207/s15327752jpa8201\\_10](https://doi.org/10.1207/s15327752jpa8201_10).
- Youngstrom, E. A. (2008). *Evidence-based strategies for the assessment of developmental psychopathology: Measuring prediction, prescription, and process*. *Psychopathology: History, diagnosis, and empirical foundations*. Hoboken, NJ: Wiley.
- Youngstrom, E. A. (2013). Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology's historical strengths to enhance utility. *Journal of Clinical Child and Adolescent Psychology*, 42(1), 139–159. Available from <https://doi.org/10.1080/15374416.2012.736358>.
- Zusne, L., & Zusne, L. (1984). *Biographical dictionary of psychology*. Westport, Conn: Greenwood Press.

# The development, expansion, and future of the WAIS-IV as a cornerstone in comprehensive cognitive assessments

4

*James A. Holdnack*

Research and Statistics Consultant, Bear, DE, United States

## Introduction

Formal measures of intelligence are a relatively new development in the course of human history, although people have compared their own abilities and aptitudes to others since the beginning of civilization. These comparisons were subjective in nature but were important in early decision-making and selection processes. Early philosophers such as Plato and Aristotle discussed reasoning and intellect, but did little to develop specific procedures to test intelligence or develop operational definitions of it. However, selection for ability to learn based on observations of performance to sample specific traits or achievements were apparent in ancient Greece, China, and the Ottoman Empire (Lindemann & Matarazzo, 1990).

In the 19th and 20th centuries a number of important developments ushered in the modern era of intelligence testing. These developments were often driven by practical needs to distinguish between individuals with different abilities levels or identify those with intellectual disabilities. One such example was the work of French psychiatrist Jean-Étienne Esquirol in 1838 to distinguish between individuals with mental illness and intellectual disability based on measurement of mental ability. Esquirol's development of methods to distinguish between these individuals, which were based on speech patterns and physical measurement, is considered by some to be the first mental test.

In England, Francis Galton's work (1869) also significantly advanced the assessment of mental abilities. He focused on measurement of general and specific abilities, as well as the development of assessment methods to measure them, although his primary approach relied on tests of sensory discrimination and motor coordination. His statistical contributions, particularly the concepts of correlation and regression to the mean, provided a foundation for modern psychometrics. Karl Pearson's work followed Galton's with the development of a number of other statistics that are commonly used in psychometric study of intelligence tests, among which included the product-moment correlation, multiple and partial correlation coefficients, and the chi-square test to assess goodness of fit. Work was also being

conducted in Germany by Hermann Ebbinghaus and Hugo Münsterberg, focusing on development of tests to assess memory, attention, learning, computation, and perception. Others working in the United States like James Cattell were developing tests of mental abilities stressing the importance of experimental study of mental abilities using laboratory-based procedures. These and other contributions during the 19th century provided a strong foundation for work to continue in intellectual assessment. Bringing to bear laboratory procedures developed in experimental psychology on the assessment of mental abilities was a critical advance for the study of individual differences as they related to mental abilities.

One notable event that has served to separate modern approaches to the assessment of intelligence from those that preceded was the publication of the 1905 Binet–Simon Scale ([Binet & Simon, 1905](#)). Alfred Binet and Theodore Simon had worked over many years in France to develop standardized methods to study mental ability and the Binet–Simon Scale represented a culmination of many of those earlier efforts. The impetus for developing the scale was at the direction of a French government commission to develop a method to identify school children with intellectual disability. The Binet–Simon scale was noteworthy because it incorporated a number of features that have persisted into current intelligence tests, including standardized instruction on how to administer test items, rank ordering items based on level of difficulty, items that assessed a number of different abilities (e.g., repeating digits, defining words, identifying similarities between words), and consideration of age associated changes in mental abilities. This relatively rudimentary scale was revised in 1908 and again in 1911 where the concept of mental age was incorporated. William Stern coined the German term intelligenzquotient, later to be translated intelligence quotient or IQ, which he suggested could be calculated by dividing the mental age from the Binet–Simon scales by the chronological age of the test participant. Binet's work was introduced in the United States by Henry Goddard in 1908 and in 1910. Working at the Vineland Training School in New Jersey, he adapted Binet's 1908 scale and standardized it on 2000 American children ([Goddard, 1910](#)). Binet's work was further advanced by Lewis Terman at Stanford University who in 1916 published a version of the Binet–Simon scale that added some additional tests, adopted the intelligence quotient proposed by Stern, and standardized the scale ([Terman, 1916](#)). Later revisions of the Binet–Simon scale with Maud Merrill in 1937 and 1960 further increased the popularity and usefulness of the Standford–Binet Intelligence Scale. The fifth edition of the Standford–Binet Intelligence Scale was released in 2003.

Group intelligence testing became popular during this time with the commencement of World War I because of the need to evaluate the intellectual functioning of military recruits. The Army alpha test was developed in 1917 for this purpose by Robert Yerkes and others, based on some of the individually administered intelligence test of Alfred Binet ([Yoakum & Yerkes, 1920](#)). Because the Army alpha test was based primarily on verbal and numerical abilities, a nonverbal equivalent called the Army beta was also developed for individuals who were illiterate or did not speak English. Test results were used to identify recruits with intellectual disabilities that would preclude them from service, identify recruits with superior abilities

who might be qualified for officer training or other special assignments, and more generally to assist the Army in assigning recruits to specific positions within the armed services. Although the Army alpha and beta are no longer used, there are a number of group intelligence tests still available that may be used when individual testing in unnecessary or impractical, such as the Multidimensional Aptitude Battery II (Jackson, 1998) and the Cognitive Abilities test (Lohman, 2011).

It was also during this time that the need for an intelligence test specifically designed for assessment of adults was noted by David Wechsler, who was critical of the Binet scales for a number of reasons. Because the Binet scales were originally developed for children, Wechsler questioned validity of the items and the concept of mental age, both which had limitations when applied to adults. He also criticized the use of a single score to reflect mental ability. Wechsler developed the first version of his adult intelligence test while chief psychologist at Bellevue Psychiatric Hospital in New York and published it in 1939 as the Wechsler–Bellevue Intelligence Scale (Wechsler, 1939). In addition to being designed for use with adults, the Wechsler–Bellevue scale differed from the Binet scales in number of other important ways. Wechsler's scale used 11 subtests organized according to content that provided scores for each of the 11 subtests as well as verbal and nonverbal (or performance) IQ scores. Organization of items based on mental age was done away with and replaced by a point-scale format originally proposed by Robert Yerkes, where credit was given for each item and the sum of the correct items became a raw score for each subtest that could then be converted to a standard score. These raw scores could also be summed to produce verbal and performance IQ standard scores based on the “deviation IQ” where 100 was the mean intelligence for the population and 15 point increments were the standard deviation. A revised version of the scale was published in 1955 as the Wechsler Adult Intelligence Scale and followed by subsequent revisions in 1981 (WAIS-R), 1997 (WAIS-III) and 2008 (WAIS-IV).

In addition to the Binet and Wechsler Scales, there are currently available other measures for individual assessment of adult intelligence including the Woodcock–Johnson Tests of Cognitive Abilities IV (Schrank, Mather, & McGrew, 2014), the Reynolds Intellectual Assessment Scales-2 (Reynolds & Kamphaus, 2015), and Raven's Progressive Matrices (Raven, Raven, & Court, 1998). Among these, the Wechsler and Binet scales have been most widely adopted. The popularity of the Wechsler scales has grown over the years since publication of the first version of the WAIS, so that by the 1960s the WAIS had become the most commonly used test of adult intelligence, and continues to be so to this day. The following sections focus on development of the most recent version of the Wechsler scales, the Wechsler Adult Intelligence Scale-IV, given that it is the most commonly administered test of intelligence. While specific to the Wechsler scales, the sections cover information relevant to current approaches to test development, measurement, and interpretation of intelligence tests. A description of the Wechsler Memory Scale—fourth edition (WMS-IV: Wechsler, 2009) and other related tests is also included as the WMS-IV was conformed with the WAIS-IV. Administration of these other tests enables clinicians to add assessment of constructs not measured by the WAIS-IV.

## WAIS-IV: development, advances, and future directions

Each revision of the Wechsler intelligence scales seeks to refine and expand the assessment of general cognitive abilities, while preserving the core measurement properties that have been validated in clinical research and practice for more than 70 years. With each revision, the zeitgeist of clinical practice affects elements of the development (e.g., length of administration, legal requirements, forensic practice, etc.) with the goal of improving functionality of the test, in addition, to supporting the strong research foundation. The Wechsler scales' broad application in school, clinical, neuro, forensic, medical–legal, and geriatric psychology are a testament to its importance in clinical practice but also illustrates the complexity of designing a tool that can meet the needs of such diverse clinical populations and practitioners. Furthermore, the Wechsler scales are used in countries throughout the world, and changes in item content, test structure, and art must be viewed with an eye toward its translation into other languages and cultures. All of these factors played a role in the development of Wechsler Adult Intelligence Scale—fourth edition (WAIS-IV: [Wechsler, 2008](#)).

## WAIS-IV development approach

The WAIS-IV development team utilized multiple methods to create a blueprint for the revision of the Wechsler Adult intelligence Scale—third edition (WAIS-III: [Wechsler, 1997a](#)). Not surprisingly, the legacy of the Wechsler Adult Intelligence Scales plays a significant role in the plan for a new edition in terms of the content of the battery (e.g., subtests) and structure (e.g., indexes and subtest to index mapping). The development team evaluates ongoing clinical research, new psychometric or theoretical approaches, and developments in neuroscience as it relates to the existing and predecessor editions to determine if changes to the content and structure should be considered for the new edition. The development team conducts survey studies of current users, field experts (e.g., cultural, linguistic, and content), researchers, and international users to identify specific item content that may be biased (e.g., contain specific art or linguistic content not known to, has a different meaning to, is offensive to, or is potentially misinterpreted by) for specific clinical or cultural groups, has instructions that are not clear to a subset of individuals or are too lengthy, or the materials are too difficult to manipulate or result in assessment of unintended abilities (e.g., fine motor abilities rather than visual-perceptual functions). The blueprint for the new edition is vetted by an advisory panel comprised of experts in clinical, school, neuro, medical–legal, clinical research, geriatrics, psychometrics, and international applications of the WAIS-III.

Surveys of WAIS-III customers identified the need for shorter administration times for the battery. In particular, a shorter test battery for older adults was cited as a significant need in the revision of the test, though most clinicians wanted the option to administer a full-battery in some cases. The primary model for reducing

testing time was to reduce the number of tests required to obtain index scores, that is, only 10 versus 13 tests are required to obtain all indexes. In addition to changing the test structure, each subtest instructions were evaluated for length and the number of items within each subtest were reviewed and shortened where possible.

Additional changes to the WAIS-III were proposed based on feedback from customers and field experts. The changes were designed to improve the assessment of the core cognitive abilities of subtest by eliminating variance attributable to other factors which can occur when evaluating individuals across a broad age range of 16–90 years. These improvements included: adding demonstration and sample items where needed, reducing vocabulary level of instructions, dropping object assembly, reducing motor demands, enlarging visual stimuli, reducing emphasis of timed performance on test not designed to evaluate processing speed, and eliminating auditory processing confounds on Digit Span and Letter–Number Sequencing. The changes not only improve the tests functionality in older adults but improves adaptability internationally and reduces nonspecific test factors that might affect certain subpopulations (e.g., lower education level).

The development team also identified areas in which the theoretical foundation of the WAIS-III could be updated based on psychometric and clinical research, theoretical models of cognitive abilities such as working memory and fluid versus visual–perceptual reasoning, and better alignment with the Wechsler Intelligence Scale for Children—fourth edition. These considerations resulted in development of new subtests Figure Weights and Visual Puzzles which are measures of fluid and visual–spatial processes and Cancellation, a measure of processing speed. Also, changes to Arithmetic and Digit Span were made to improve assessment of the core working memory construct. Finally, the dual IQ model (e.g., Verbal and Performance) from the WAIS-III was eliminated.

The WAIS-III subtests were reviewed for dated items and overall psychometric properties. New items were developed to replace items that were dropped. Additionally, all verbal responses were reevaluated for 0, 1, 2 scoring rules as the nature and type of responses change as a function of changes in the culture (e.g., impact of technology) and education level of the population. All subtests were evaluated for floor and ceiling effects. A particular emphasis was made to improve reliability of assessment at the critical cutoff score of 70. In order to achieve these goals, new items at the floor and ceiling levels were created and evaluated. While item and subtest level changes were an important consideration in the development of the WAIS-IV blueprint, the theoretical and clinical interpretability of the battery were of central importance.

In a continuation of the evolution of the Wechsler model, the WAIS-IV index scores are the primary focus of clinical interpretation. This is a shift from the original Wechsler model of intelligence which was based on a two-part structure comprised of the Verbal IQ (VIQ) and Performance IQ (PIQ). The current Wechsler factor model includes four factors: Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed. The shift in focus to more specific factor-based indexes began in 1991 when the WISC-III introduced four factor-based index scores as an optional alternative to the traditional VIQ/PIQ structure; the WAIS-III

followed suit in 1997. In 2003, the WISC-IV dropped VIQ and PIQ, eliminating them from the test in favor of the four factor-based structure presented in WISC-III. In addition, the name of the Freedom from Distractibility Index was changed to the Working Memory Index to reflect the increasing understanding of the construct, the name of the Perceptual Organization Index was changed to the Perceptual Reasoning Index to accurately describe the increased focus on reasoning skills among the newly created visual—perceptual subtests, and clinical interpretation was focused on the four index scores. The WAIS-IV adopted the four-factor model in 2008 and further refinement to the Wechsler model occurred with the publication of the WISC-V in 2013. The WISC-V now includes five factors: Verbal Comprehension Index, Visual—Spatial Index, Fluid-Reasoning Index, Working Memory Index, and Processing Speed, illustrating the continued evolution of the Wechsler Scales.

## **Subtest level changes**

Subtest changes were implemented to achieve the goals of improving the functionality and psychometric properties of subtests. One of the primary goals of the revision was to reduce the impact of nonconstruct related variance. To this end, Picture Arrangement and Object Assembly were dropped as subtests from the third to fourth edition. These tests require intact fine motor coordination and are also among the more difficult tests to administer (e.g., require a lot of materials and coordination by the examiner). Of the subtests that remained from the third edition, small changes to item content, administration rules, and scoring were made to improve the psychometric properties, remove outdated content, and insure examinees understand how to perform the test. Subtest level changes are presented by construct in the subsequent sections.

### ***Verbal Comprehension***

The primary changes to the Verbal Comprehension subtests were in the administration rules, scoring, and item content changes. Scoring studies were performed on the standardization sample to identify any changes required in scoring items that remained from the third edition. These changes often reflected the impact of technology and other cultural changes occurring since the last standardization. For new items, scoring rules were developed based on the responses of individuals with low, medium, and high ability. The psychometric impact of each response score was closely evaluated. The overall stimulus structure remained the same for all the subtests with the exception of the addition of picture naming items which were added to improve the assessment of very low functioning examinees. Discontinue rules were change for Vocabulary and Information, going from 6 on the third edition to 3 on the fourth edition. Similarities and Comprehension discontinue rules were reduced from 4 to 3. These changes help reduce administration time while not

significantly impacting the psychometric properties of the test. Item Response Theory (IRT) was employed to identify the optimal number of items in each subtest and to identify poorly fitting items.

### ***Perceptual reasoning***

The Block Design subtest administration procedures were slightly modified to include the use of both the model and the stimulus book for all teaching trials, which better illustrate to the examinee that they would be making the same design as shown in the stimulus book. The number of items with time bonuses was reduced to lessen the impact of speed of performance. This reduction was possible with the addition of diamond-shaped items prior to the 9-block administration. These items improved the item-difficulty gradient and reduced the need to rely on bonus points. The impact of speed can be further reduced by using the supplemental Block Design no bonus time score. In the third edition, Matrix Reasoning contained four types of items; however, examinees were not trained on all the item types. For the fourth edition, Matrix Reasoning contains only two item types and each type of item is trained at the beginning of the test to insure the test is measuring reasoning and not a misunderstanding of a change in item type. Few changes were made to Picture Completion. These entailed enlarging the stimuli and addition of more difficult items to improve the test ceiling. Discontinue rules were reduced for Block Design (3–2), Matrix Reasoning (4–3), and Picture Completion (5–4). IRT was used to evaluate optimal number of items for each subtest and to identify poorly fitting items.

### ***Working Memory***

Significant changes were made to the Working Memory subtests with the intention of improving assessment of the core construct and reducing the impact of nonconstruct related variance. The Digit Span subtest includes a new number sequencing condition in addition to the traditional forward and backward conditions. Rhyming numbers were eliminated within a trial on Digit Span. Arithmetic items containing references to the English measurement system were dropped or altered to eliminate that specific type of knowledge from affecting performance on the test. This change improves the purity of the measure and increases portability to other countries. The math skills required to complete Arithmetic items was significantly reduced and the working memory demands increased by requiring multiple steps rather than more complex computations. The discontinue rule was reduced from 4 to 3 for Arithmetic. Like Digit-Span, rhyming letters and numbers were eliminated in Letter–Number Sequencing trials. Also, a graduated training procedure was implemented on Letter–Number Sequencing.

### ***Processing Speed***

The Processing Speed subtests were modified to reduce the impact of nonconstruct related variance. On Symbol Search, the stimuli were increased and the response

process changed. In this edition of the test, examinees do not mark the ‘yes’ or ‘no’ box but must cross-out the symbol if it is present or mark ‘no’ if it not present. This insures that the examinee actually got the yes item correct and did not guess it correctly. The Coding test was minimally changed, however, the stimulus presentation order was modified to have equal exposure to all symbols within a row. Slight changes were made to the administration to be more directly aligned with the WISC-IV.

### ***New subtests***

The elimination of the Picture Arrangement and Object Assembly subtest created a need for new perceptual reasoning subtests to be developed. With primary goals of reducing the impact of speed and fine-motor contribution to visual–perceptual measures and with a desire to increase the level of fluid reasoning assessed, two new subtests were created. A third new subtest was created in the Processing Speed domain, to allow for substitution when one of the core measures is spoiled and to parallel the WISC-IV processing speed measures. The new subtests are Visual Puzzles, Figure Weights, and Cancellation.

### ***Visual Puzzles***

The Visual Puzzles subtest was designed to replace the visual construction task, Object Assembly. On the Visual Puzzles task, the examinee sees a design at the top of the page and they must select the three designs that could be pieced together to make the design at the top of the page. The examinee must mentally match global and local features and in some cases mentally rotate pieces to determine if they would be the best fit. The test is a more pure measure of visual–spatial processing as it does not have bonuses for speeded performance nor does it require physical manipulation of any puzzles pieces.

### ***Figure Weights***

The Figure Weights test was not developed as a direct nonmotor analog of the Picture Arrangement subtest. Rather it was designed to increase the level of fluid reasoning in the Perceptual-Reasoning domain, particularly quantitative reasoning. On the Figure Weights test, the examinee is shown a scale with different shapes on one side of a scale and a second set of shapes that have the equivalent weight as the first set of shapes. The examinee must determine how the shapes are related (e.g., one square equals two triangles). Once they determine the relationship between the shapes, they must apply that knowledge to a second scale to balance the weight. Like Matrix Reasoning, the examinee selects the correct answer from five possible choices. This test does not use time bonuses and no physical materials are manipulated.

## Cancellation

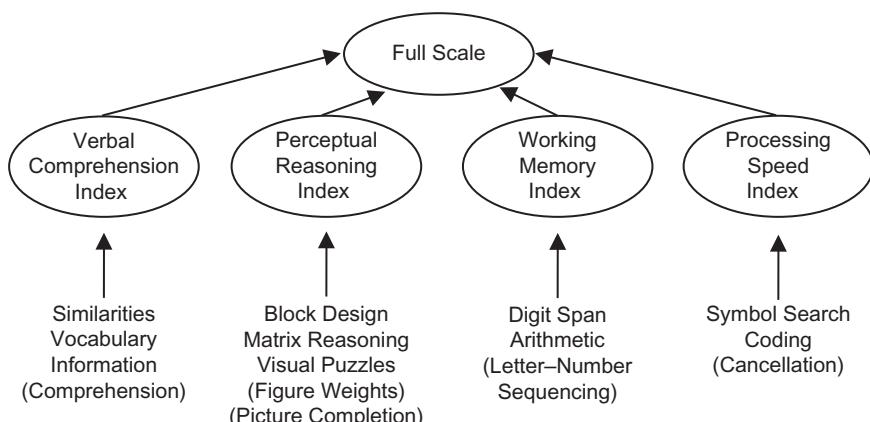
The Cancellation subtest measures visual scanning much like its counterpart on the Wechsler Intelligence Scale for Children—fourth edition (WISC-IV: [Wechsler, 2003](#)). The design of the WAIS-IV version differs from the WISC-IV version in that it does not use common objects that must be identified. Rather, the examinee must find all the shapes that match the targets in shape and color. The use of two dimensions creates an inhibitory effect in which the shape may be correct but the examinee must inhibit the response because the color is incorrect. The cancellation test is timed and is primarily a measure of visual processing speed.

## WAIS-IV index scores and structure

[Fig. 4.1](#) presents a visual model of the WAIS-IV. The composite scores, comprised of the four index scores and FSIQ, are presented in ovals and are the primary focus of interpretation. The subtests are listed below the index in which they are included. Parentheses are used to differentiate supplemental subtests from the core subtests used to derive index scores.

### Verbal Comprehension Index

The Verbal Comprehension Index (VCI) reflects an individual's ability to comprehend verbal stimuli, reason with semantic material, and communicate thoughts and ideas with words. As shown in [Fig. 4.1](#), the VCI consists of the Similarities (SI), Vocabulary (VC), and Information (IN) subtests. The Comprehension (CO) subtest is supplemental. All VCI subtests are well-known and extensively researched measures. The WAIS-IV VCI correlates  $r = .91$  with the WAIS-III VCI showing a very high degree of concordance.



**Figure 4.1** WAIS-IV factor model.

## ***Vocabulary***

The Vocabulary subtest assess word knowledge and the ability to access and to effectively communicate that knowledge. The Vocabulary subtest is influenced by a number of cognitive processes including expressive and receptive language abilities; declarative, long-term memory; and working memory (e.g., relates to memory search). Observation of responses can provide insight into the cognitive processes invoked for each individual. Individuals with a long-history of language difficulties will have a lack of content knowledge but may also have difficulties with verbal expression. Individuals with memory difficulties may respond with near misses or circumlocution. Those with deficits in working memory functioning may give rapid and long responses of poorly articulated information due to a failure to evaluate information for accuracy and efficiency prior to verbal expression. Vocabulary has a high correlation across versions ( $r = .87$ ).

## ***Similarities***

The Similarities subtest assesses word knowledge, abstract reasoning, and the ability to effectively communicate verbal information. The expressive language demands tend to be less than those for Vocabulary. Similarities requires a different form of word knowledge. Most of the stimulus words are well known; however, the solution to the problem often requires knowledge of secondary or tertiary meanings of the same words. Also, it requires the distillation of the meaning down to factors that can be related across the words. In this way, the test requires some degree of cognitive flexibility and abstract reasoning, in that, if the person can only think of a single meaning to a word, they may get stuck in their responding. Very concrete individuals may give responses that reflect only superficial aspects of the words (e.g., “have skin” for limbs). Individuals with poor working memory skills may give a number of disjointed responses until they hit upon an answer. Similarities has a high correlation between versions ( $r = .75$ ).

## ***Information***

Information evaluates the examinees general fund of knowledge and the ability to access and express this information. This test places high demands on long-term, declarative memory functions; and working memory skills to facilitate search on more difficult items. Information requires less language processing compared to other Verbal Comprehension measures. Reasoning can influence performance on Information as examinees may identify possible answers and go through a process of reasoning as to which is the most probable correct answer. Information has a very high correlation between versions ( $r = .90$ ).

## ***Comprehension***

Comprehension is the most cognitively complex of all the Verbal Comprehension measures. It requires the ability to access information but in a manner that requires

more reasoning than Vocabulary or Information. The linguistic demands, both receptive and expressive, are greater than on other tests in the domain. Comprehension also requires a degree of cognitive flexibility due to the requirement to make multiple responses to a single item. Working memory and memory skills are needed to access and sort through competing information, and to formulate a response to the question. Comprehension pulls for very long responses in individuals with reasoning difficulties or working memory issues. Comprehension has a high correlation between versions ( $r = .74$ ).

### ***Perceptual Reasoning Index***

The Perceptual Reasoning Index (PRI) measures fluid reasoning and visual-perceptual processing. The core WAIS-IV PRI subtests are Block Design (BD), Matrix Reasoning (MR), and Visual Puzzles (VP). The Figure Weights (FW) subtest is supplemental. Block Design and Matrix Reasoning are well known measures of visual-spatial organization and fluid reasoning, respectively. Picture Completion (PC) is a cognitively complex measure of visual perceptual and reasoning abilities. In the WAIS-III, this index was called Perceptual Organization and the change in the name reflects the mix of perceptual and reasoning abilities that comprises the index. The two versions correlate  $r = .84$  indicating a high degree of concordance between the measures.

There are two subdomains within the perceptual reasoning index. A visual-spatial domain and a fluid reasoning domain. Block Design and Visual Puzzles measure aspects of visual-spatial processing. These tests focus on *constructional* abilities. That is, the use of visual information to build a geometric design to match a model. This is a reasoning task that involves the ability to identify the spatial relationships and visual details of objects for the purpose of building a new design. The parts must be seen as elements of the whole design. It is a reasoning task because the solutions require more than simply matching a part to a part in the design. Mental rotation and visualization of the solution is required.

The fluid reasoning subdomain is assessed by Matrix Reasoning and Figure Weights. These tests use visual information to identify a common theme or concept which can often be verbalized. The visual information does not directly provide a solution to the problem. Rather, the relationship among visual-spatial elements provides clues as to the single underlying concept that binds them all together. Once the examinee figures out the underlying conceptual link, they must be able to apply that knowledge to identify the correct solution.

### ***Visual Puzzles***

This subtest measures mental, nonmotor, construction ability which requires visual and spatial reasoning, mental rotation, visual working memory, understanding part-whole relationships, and the ability to analyze and synthesize abstract visual stimuli. Deficits in any aspect of visual processing may impact performance on this test.

### ***Block Design***

Block Design measures ability to construct a design to a model. It requires visual–spatial processing and integration. There are less demands on visual working memory compared to Visual Puzzles as the examinee does not need to imagine how connecting parts together would appear visually. There is a procedural and active learning component to Block Design as the same stimuli are used across items (e.g., once figure out how to make a stripe can use the knowledge on later items). The use of physical stimuli provides the examinee with immediate, concrete, visual feedback on correctness of solutions or partial solutions and in some cases a trial and error approach will stumble on a correct partial solution. Observationally, Block Design provides a rich source of understanding the examinee approach to solving visual–construction problems. The test does require adequate processing speed and fine-motor skills. Block Design has a moderate correlation between editions ( $r = .77$ ).

### ***Matrix Reasoning***

Matrix Reasoning requires the examinee to observe changes in visual information and to link the changes in visual information to an underlying concept. The visual–spatial information provides clues to how the objects are related. Once the examinee understands the relationships among the objects they can apply that rule to find the correct answer which is the object that follows the same rules as the test stimuli. Severe deficits in visual-perceptual problems can impact results. The test requires adequate working memory as many solutions often need to be considered to identifying the underlying construct. Matrix Reasoning has a moderate correlation between versions ( $r = .71$ ).

### ***Figure Weights***

Figure Weights requires the examinee to apply the quantitative concept of equality to understand the relationship among objects and then use the concepts of matching, addition, and/or multiplication to identify the correct response. Like Matrix Reasoning, visual information informs the child about the conceptual relationship among objects but in this case the concept relates to how they are equivalent. Unlike Matrix Reasoning, the child does not have to follow a series of steps to identify the linking concept. Rather, the difficulty lies in applying the equality rule and selecting the proper quantitative operation to get the correct response. The test requires some knowledge of math and individuals with severe impairments in adding and/or multiplying may find the task difficult.

### ***Picture Completion***

Picture Completion requires the identification of visual clues that enable the examinee to reason as to the underlying “missing element” that is to be identified. In this respect, this subtest is cognitively similar to Matrix Reasoning and Figure Weights,

in that, the visual clues provide insight into the underlying theme which can often be verbalized. Picture Completion requires visual scanning, visual detail/local processing, conceptual reasoning (e.g., what should be present), global versus local processing, and confrontation naming skills (e.g., most examinee name missing element). Picture Completion has a lower concordance across versions than other measures ( $r = .65$ ).

## ***Working Memory Index***

The Working Memory Index (WMI) measures attention, concentration, and working memory. The WMI consists of two primary subtests, Digit Span and Arithmetic, and the supplemental Letter–Number Sequencing. The Digit Span subtest includes three tasks: Digit Span Forward (DSF), Digit Span Backward (DSB), and the new Digit Span Sequencing (DSS). DSF measures short-term memory, not working memory. DSB and DSS measure auditory working memory. The current reformulation of the Arithmetic subtest represents a substantial transformation from previous editions and is a much improved measure of working memory. Compared to its predecessors, the WAIS-IV Arithmetic subtest contains reduced verbiage in the word problems, fewer items with time bonuses, and simpler numerical calculations with more steps, reducing the pull of mathematical skill and increasing the activation of working memory. The WMI correlates  $r = .87$  across versions indicating a high degree of concordance between the two editions.

### ***Digit Span***

While Digit Span is comprised of three distinct conditions, there are a number of cognitive processes that impact performance across all the conditions. Each of the conditions requires intact auditory processing/discrimination, focused attention, registration, and sustained effort. The phonological loop may be activated in all conditions, though early items on forward and backward trials may be completed rotely. Some examinees will employ executive functions in the form of strategic recall (e.g., clustering) on difficult items across conditions. The forward condition assesses span capacity. The backward condition primarily evaluates the ability to manipulate information in working memory. In addition to mental manipulation, the backward condition may also be affected by procedural learning (e.g., exposure to forward condition helps examinee understand the nature of the task). The backward condition requires a fixed response order such that verbal responses can be initiated and rehearsed as soon as the first digit is presented. The sequencing condition requires a more active form of mental manipulation as numeric sequencing requires a constant comparison of elements held in storage. Additional skills such as knowledge of the value of numbers on the number line and ability to track repeated numbers are also required. The verbal response cannot be rehearsed until the last digit is presented. Procedural learning may influence performance as the examinee has had two prior exposures to the test format. Digit Span has a moderate correlation between test editions ( $r = .77$ ).

### ***Arithmetic***

Arithmetic is a complex measure requiring a number of cognitive skills of which working memory in the form of phonological loop and mental manipulation are the primary components. Like Digit Span Sequencing, Arithmetic requires an active form of mental manipulation. Elements are held in storage until they can be combined and transformed into the correct answer. The verbal response cannot be developed until the last piece of information is obtained. In addition to working memory, the Arithmetic subtest requires quantitative reasoning, computational abilities, and verbal/auditory processing. In factor analytic studies, this subtest frequently cross-loads on multiple cognitive domains. Arithmetic has a moderate correlation between test versions ( $r = .79$ ).

### ***Letter–Number Sequencing***

The Letter–Number Sequencing subtest assesses active mental manipulation, similar to the Sequencing condition of Digit Span, and dual-tasking (e.g., sorting and holding two types of stimuli in working memory). Each of the elements must be compared and sorted and the verbal response sequence cannot be rehearsed until the last piece of information is obtained. Like Digit Span, Letter–Number Sequencing requires auditory discrimination, focused attention, registration, and effort. Letter–Number Sequencing uniquely requires alphabetic knowledge which might affect performance in individuals having weak English skills or a history significant learning problems. The correlation between Letter–Number Sequencing across versions is high at  $r = .70$ .

### ***Processing Speed Index***

The Processing Speed Index (PSI) measures the speed of mental processing, using visual stimuli and graphomotor skills, and is related to the efficient use of other cognitive abilities. One of the primary elements of all the processing speed tasks is speed of decision-making. While the ability to rapidly identify and discriminate visual information is important, the ability to decide if an answer is correct and to implement the response is a critical component. This can be observed in individuals with a high degree of anxiety, where the response is slowed by uncertainty and not necessarily slow processing of visual information. This should not be considered an unintended or nonconstruct related form of variance as slowed decision-making has a direct impact on daily functioning. Automaticity is an important construct in evaluating processing speed measures. Some measures, like naming tasks have a very high degree of automaticity, such that tasks requiring identification of overlearned information can be completed with little thought. High automaticity yields very fast responding in healthy adults, whereas, processing speed tasks with low automaticity will result in slower more deliberate and more variable performance.

The processing speed subtests are less cognitively complex than most of the WAIS-IV measures; however, performance can be affected by a number of

construct and nonconstruct related factors. The PSI includes the core Coding and Symbol Search subtests and the supplemental Cancellation subtest. Cancellation is similar to previously developed cancellation tasks designed to measure processing speed, visual selective attention, vigilance, perceptual speed, and visual-motor ability (Bate, Mathias, & Crawford, 2001; Geldmacher, Fritsch, & Riedel, 2000; Sattler & Ryan, 2009). Cancellation tasks have been used extensively in neuropsychological settings as measures of visual neglect, response inhibition, and motor perseveration (Adair, Na, Schwartz, & Heilman, 1998; Geldmacher et al., 2000; Lezak, Howieson, & Loring, 2004; Na, Adair, Kang, Chung, Lee, & Heilman, 1999). Relative to the WISC-IV version of Cancellation, an inhibitory component was added in the WAIS-IV to place more complex cognitive demands on examinees. The examinee must simultaneously discriminate both the color and the shape of the stimuli and inhibit responding when only one of the two features is present. PSI has a high correlation between test editions ( $r = .86$ ).

### ***Coding***

The Coding subtest primarily measures ability to rapidly pair visual stimuli, and rapid decision-making about the response. The task has low levels of automaticity, at least initially, because the knowledge of numbers does not present any advantage to knowing the correct response. However, due to the effects of procedural learning the task may become more automatic on later items (e.g., reduces the burden on decision-making). In addition to processing speed, Coding requires attention to detail, visual discrimination, incidental learning (paired associate), a moderate degree of fine-motor control, and ability to draw/construct designs from a model. Coding was previously referred to as Digit-Symbol Coding on the WAIS-III. The name was changed for consistency with the WISC-IV. The correlation between Coding and Digit-Symbol Coding is high at  $r = .85$ .

### ***Symbol Search***

Symbol Search assesses rapid visual scanning and identification and decision-making. Symbol Search has a higher degree of automaticity, as simple visual matching, while not relying on prior knowledge, is a highly entrenched ability. Items containing a matching response will not evoke a decision reaction as much as items without a matching symbol, that is, once a match is found most examinees will move on to the next item, though more anxious examinees may double-check if the response is correct. “No” response items tend to pull for more checking behavior “Did I miss the answer?” and pull for more decision-making about moving onto the next item. Symbol Search has less fine-motor demands compared to Coding. Symbol Search has a high correlation between the third and fourth editions ( $r = .72$ ).

### ***Cancellation***

Cancellation has the highest degree of automaticity among the WAIS-IV processing speed measures. The ability to rapidly identify shapes and colors develops from a

very early age. In the case of Cancellation, the high degree of automaticity results in very rapid responding, which in some cases can produce erroneous responses, as stimuli matching on one feature but not both features will “pull” for a response. These inhibitory elements increase the decision-making load of the test. The test primarily measures visual scanning and rapid visual identification.

### **General Ability Index**

New to the WAIS-IV, the General Ability Index (GAI) and Cognitive Proficiency Index (CPI) are two optional index scores that may be calculated to reflect performance across the four index scores. The General Ability Index (GAI) summarizes performance on the VCI and PRI in a single number. These two indexes are traditionally thought to contain the most highly ‘g’ loaded subtests within WAIS-IV. The WAIS-IV GAI excludes the contributions of the WMI and PSI to intelligence because the latter measures are frequently affected in clinical conditions. Clearly, GAI and FSIQ can lead to different impressions of a patient’s overall ability when there is variability across the four indexes.

The GAI was first developed for use with WISC-III in ability–achievement discrepancy analyses (Prifitera, Weiss, & Saklofske, 1998) because many students with a learning disability (LD) exhibit cognitive processing deficits in working memory and processing speed concomitant with their learning disabilities. This pattern can result in lower FSIQ scores for individuals with LD, and consequently smaller discrepancies with achievement, thus making it less likely that he or she will be identified as underachieving and in need of services. Thus, use of GAI as a measure of global intelligence for comparisons with other abilities impacted by lower performance on working memory and processing speed, such as achievement, is often a more appropriate comparison with those constructs than the FSIQ.

Other uses for the GAI have since been identified. For example, the GAI may be an appropriate estimate of overall ability when physical or sensory disorders invalidate performance on the working memory or processing speed tasks, or both. Another possible use is in approximating the pre-injury cognitive status and memory abilities of patients with traumatic brain injury. All discrepancies between the WAIS-IV and WMS-IV utilize the GAI in place of the FSIQ. Memory impairment is often compromised in the same clinical conditions where low WMI and PSI scores are observed (e.g., moderate-to-severe TBI). Similar to the differences observed in LD evaluations, when FSIQ is lower due to comparatively low WMI and PSI scores, it is more difficult to find differences between intellectual functioning and memory functioning. The VCI and PRI tend to be less affected in clinical disorders in which memory functioning is also impacted. The use of GAI allows the practitioner to better identify memory deficits in patients with concomitant PSI and/or WMI deficits.

Prifitera, Saklofske, and Weiss (2005) suggest that some practitioners may prefer the GAI as an alternative way of summarizing overall ability. This suggestion has led to an increasing number of psychological evaluations in which the GAI is described as a better estimate of overall ability than FSIQ whenever the WMI or

PSI scores are significantly lower than the VCI or PRI scores. As subsequently clarified, this is not what the authors intended (Prifitera, Saklofske, & Weiss, 2008). The GAI should be considered a better estimate of intelligence than FSIQ only when there are sound clinical reasons to exclude WMI and PSI, such as invalid administration due to lack of effort, sensory or physical impairments, or disturbance of the testing session. Working memory and processing speed are essential components of a comprehensive assessment of intelligence, and excluding them from the estimate of overall intelligence simply because the patient's abilities in those areas are relative weaknesses is poor practice. Such practice will result in unrealistically high estimates of intelligence for these patients, possibly excluding them from needed services and creating unrealistic employment expectations. The GAI can be a vital comparison measure in an evaluation when described correctly; however, it is not the best estimate of overall intelligence because it suffers from construct underrepresentation.

### **Cognitive Proficiency Index**

The Cognitive Proficiency Index (CPI) summarizes performance on the working memory and processing speed indices of the WAIS-IV in a single score. The CPI represents a set of functions whose common element is the proficiency with which one processes certain types of cognitive information. Proficient processing, through quick visual speed and good mental control, facilitates fluid reasoning and the acquisition of new material by reducing the cognitive demands of novel or higher order tasks (Weiss, Saklofske, Prifitera, & Holdnack, 2006). In other words, efficient cognitive processing facilitates learning and problem-solving by "freeing up" cognitive resources for more advanced, higher level skills.

The WAIS-IV CPI excludes the contributions of verbal comprehension and perceptual reasoning to intelligence. Thus, CPI and GAI can provide different views into a patient's cognitive abilities when there is significant variability across the relevant index scores. Both views are sometimes necessary to form a complete picture of an individual's strengths and weaknesses that is not distorted by combining a set of diverse abilities into a single overall score. Rather than reporting GAI as the best estimate of overall ability when the profile of abilities is diverse, it is sometimes better practice to describe both the GAI and CPI in the psychological evaluation. Normative tables for the WAIS-IV CPI are provided in Weiss, Saklofske, Coalson, and Raiford (2010).

### **Issues in summarizing overall ability**

The FSIQ has strong explanatory and predictive power at the group and individual level. Still, the use of an overall summary score may mask individual differences among the broad domains of general ability, especially in patients with neuropsychological deficits where the focus of clinical attention is not prediction but

diagnosis of underlying cognitive deficits. For this reason, the relevance of reporting IQ scores has been questioned (Fiorello et al., 2007). Yet, other researchers suggest that FSIQ may be an equally valid measure of general ability for individuals or groups having highly variable index scores as for those having consistent index scores (Daniel, 2007) and that there may be no difference in the predictive validity of FSIQ for low-scatter and high-scatter groups (Watkins, Glutting, & Lei, 2007).

FSIQ is an especially strong predictor of educational attainment, occupational level, memory functioning, and school achievement (Wechsler, 2008, 2009). FSIQ and achievement, for example, correlate strongly, typically around  $r = .70$ . This means that FSIQ explains about half the variance in achievement. There is no known variable or collection of variables that come close to fully accounting for the other half. Beyond the relationship with achievement, there is considerable ecological and criterion validity for the use of an overall estimate of general intelligence in a variety of areas related to success in life including preemployment testing and predicting job performance (Gottfredson, 1997, 1998; Kuncel, Hezlett, & Ones, 2004). Thus, when the focus of clinical attention is prediction then FSIQ is often the most potent predictor. When the focus is clinical diagnosis of pathology then the index scores are often more informative. In addition, FSIQ has a very high concordance across versions of the test ( $r = .94$ ) indicating that it is the most stable estimate of ability of all the index and subtest measures.

## Five-factor models

The recently published Wechsler Intelligence Scale for Children—fifth edition (WISC-V: Wechsler, 2013) expands the index structure of the Wechsler scales from four factors to five factors. In the WISC-V, the index structure includes: Verbal Comprehension, Fluid Reasoning, Visual–Spatial, Working Memory, and Processing Speed. The addition of Visual Puzzles and Figure Weights enabled the development of five versus four indexes. In addition to the new core Fluid Reasoning Index, the WISC-V also offers a supplemental Quantitative Reasoning Index comprised of Figure Weights and Arithmetic. Prior to the publication of the WISC-V, research of the WAIS-IV factor structure suggested the possibility of a fifth factor based on the WAIS-IV subtests.

Benson, Hulac, and Kranzler (2010) argued for a five-factor model in which Matrix Reasoning, Figure Weights, and Arithmetic loaded on a Fluid Reasoning factor. Ward, Bergman, and Herbert (2011) proposed a modification of the WAIS-IV four-factor model to include Visual–Spatial Reasoning (Block Design, Visual Puzzles, and Picture Completion) and Quantitative Reasoning (Figure Weights and Arithmetic) factors. Weiss, Keith, Chen, and Zhu (2013) tested a five-factor model in which Quantitative Reasoning (Arithmetic and Figure Weights) was defined as a narrow ability subsumed under the Fluid Reasoning factor (Matrix Reasoning, Figure Weights, and Arithmetic). The above studies included the supplemental subtests, which is clearly best practice for model definition but often of limited utility to

practitioners who have administered only the core subtests. Using only core WAIS-IV subtests, [Lichtenberger and Kaufman \(2009\)](#) proposed a five-factor model in which the fluid factor consists of Matrix Reasoning and Arithmetic (without Figure Weights which is supplemental).

These WAIS-IV studies should be interpreted in light of the literature on the WISC-IV factor structure because the two tests have the same conceptual model, albeit with a couple different subtests. Extant research on WISC-IV factor structure contains a similar set of findings as WAIS-IV: both four- and five-factor models emerge as good fits to the data and the fifth factor is often characterized as fluid reasoning, although complicated by the interpretation of Arithmetic as fluid or quantitative reasoning ([Bodin, Pardini, Burns, & Stevens, 2009](#); [Chen, Keith, Chen, & Chang, 2009](#); [Chen, Keith, Weiss, Zhu, & Li, 2010](#); [Chen & Zhu, 2012](#); [Keith, Fine, Taub, Reynolds, & Kranzler, 2006](#); [Lecerf, Rossier, Favez, Reverte, & Coleaux, 2010](#); [Weiss et al., 2013](#)).

Based on consistent evidence that the WISC-IV and WAIS-IV factor structure may include a fifth factor and confirmation of five factors in the WISC-V, there is strong support for the presence of a fifth factor within the WAIS-IV. Furthermore, a theoretical quantitative factor comprised of Figure Weights and Arithmetic has been explored in the WISC-V and in a combined WAIS-IV/WMS-IV quantitative factor ([Holdnack, Drozdick, Weiss, & Iverson, 2013](#)).

## WAIS-IV and digital assessment

The Wechsler Adult Intelligence Scale—fourth edition was one of the first cognitive batteries to be adapted for digital administration on the Q-interactive assessment and reporting platform. The Q-interactive platform uses two iPads tethered through a Bluetooth connection for the administration of cognitive tests. One iPad provides all the administration information required for an individual subtest. The examiner uses the iPad to read instructions, complete sample and training items, score responses, present stimuli, and time responses. The iPad tracks the examinees performance to inform the examiner about start, reversal, and discontinue rules, thus relieving the burden of tracking these administrative tasks. The examiner may override the administration rules if they wish to test limits with a specific examinee. Subtests are immediately scored and transformed into scaled scores allowing the examiner to track the examinees performance in real-time and make determination if additional tests are needed. Additional subtests from the WAIS-IV or from other test batteries can be added on-the-fly if the examiner wishes to test additional cognitive abilities and clinical hypotheses.

The second iPad serves as the stimulus book for presenting items to the examinee. The item timing is controlled by the examiner such that the examiner does not need a stopwatch to know when to change stimuli. The examinee responds in the same exact manner as in the paper and pencil version of the WAIS-IV. For verbal subtests, the examinee simply responds to the questions and the examiner records

and scores the response on the examiners iPad. The examiner can write verbatim responses and write notes in a manner similar to the way in which they can on the paper form. Also the examiner can choose to record verbal responses for scoring at a later time. For Block Design and Picture Completion, the examinee sees the stimuli on the screen but respond in the manner used in the paper edition. Similarly, Coding, Symbol Search, and Cancellation are all completed on the paper response sheets used during the standardization of the paper form, to insure equivalence in response processes. Working memory subtests, Digit Span, Arithmetic, and Letter–Number Sequencing are administered in exactly the same way as in the paper version, with only Arithmetic requiring stimulus presentation on the iPad. The only subtests that allow an examinee to respond directly on the iPad are Matrix Reasoning, Figure Weights, and Visual Puzzles. These multiple choice response tests lend themselves to a simple touch on the screen. The examiner may override a response if the examinee states they touched the wrong one or the examiner may directly enter the response if the examinee is not able to use the touch screen.

Q-interactive was designed to give examiners and examinees the same experience as a paper and pencil administration, with the bonus of freeing examiners from mundane administrative tasks. The computer-guided administration should improve reliability of assessment and increase the examiner's ability to make performance observations. Unlike typical computer-administered tests, the examiner has full control over the test session and test administration, as the computer only supports their ability to perform routine aspects of testing. An equivalence study (Daniel, 2012) between paper and pencil and digital administration was completed as part of the digital adaptation of the WAIS-IV. The results demonstrated very consistent outcomes obtained between paper and pencil and digital administration. There were some specific effects related to faster times on processing speed and lower scores on verbal comprehension, but these have minimal clinical implications (Daniel, 2012).

## **Expanded assessment**

The WAIS-IV continues the comprehensive assessment model established with WAIS-III. The conorming of the WAIS-IV with the Wechsler Memory Scale—fourth edition (WMS-IV: Wechsler, 2009) and the Advanced Clinical Solutions for the WAIS-IV/WMS-IV (ACS: Pearson, 2009) enables clinicians to add assessment of constructs not measured by the WAIS-IV and refine the analysis of WAIS-IV results to customize the assessments to answer a variety of clinical questions. Direct comparisons between the WAIS-IV and other cognitive measures identifies if memory or social perception deficits are present beyond general low cognitive functioning or due to specific cognitive processes evaluated on the WAIS-IV. Using tools from the ACS, the clinician can determine if WAIS-IV scores represent a decline in cognitive functioning considering the background and literacy level of the examinee, or if suboptimal effort may be lowering WAIS-IV scores unexpectedly.

## **WMS-IV**

The Wechsler Memory Scale—fourth edition (Wechsler, 2009) is the most recent revision of the Wechsler Memory Scale—third edition (Wechsler, 1997b), and its development timeline paralleled that of the WAIS-IV. The two batteries were intentionally developed at the same time for the purposes of coordinating content coverage and to conorm the tests at standardization. The conorming of these two batteries is not novel, as the tests were conormed for WAIS-III/WMS-III, as well. The novelty in this revision is the intentional focus on having complimentary measures with no content overlap between the batteries. In the WAIS-III/WMS-III, the Working Memory Index shared Letter–Number Sequencing and Digit Span across the batteries but the other subtest contributors to the indexes were different. In the revision of the two batteries, it was decided to eliminate the overlap and focus on making complimentary working memory measures. The WMS-IV is comprised of five primary index scores; Visual Working Memory, Auditory Memory, Visual Memory, Immediate Memory, and Delayed Memory. In addition to core memory measures, the WMS-IV has a cognitive screening subtest to identify severe cognitive impairment.

### ***Visual Working Memory***

The Visual Working Memory Index (VWMI) was created to be a visual analog of WAIS-IV auditory working memory measures. The VWMI is composed of the Spatial Addition and Symbol Span subtests, neither of which appear in the WAIS-IV nor the WMS-III. The VWMI measures the ability to temporarily hold and manipulate spatial locations and visual details. Spatial and visual detail-based subtests were developed to account for the two visual pathways (e.g., dorsal and ventral) that process visual information in the brain. Both subtests require mental manipulation of information in working memory which was not well assessed in the predecessor visual working memory measure on the WMS-III (i.e., Spatial Span). While a low score on the VWMI indicates difficulties with visual working memory functioning, other cognitive problems may influence performance and need to be considered when interpreting results including visual discrimination problems, visual–spatial processing impairment, severe attention problems, and impaired executive functioning.

Performance on the WAIS-IV WMI can be compared to the WMS-IV VWMI Index by simple difference method or using a contrast scaled score (i.e., Working Memory vs Visual Working Memory Contrast Scaled Score). The comparison between these two measures indicates if there is a modality specific deficit in working memory (e.g., auditory vs visual). The two indexes are moderately correlated with one another ( $r = .62$ ). The WAIS-IV PRI can be compared to the VWMI index to determine if visual–perceptual deficits are reducing VWMI performance or vice versa. Low scores indicate unexpectedly low VWMI scores controlling for perceptual reasoning abilities. High scores suggest PRI is lower than expected compared to visual working memory skills. The two indexes have a moderate correlation ( $r = .66$ ).

### *Spatial Addition*

For Spatial Addition, the examinee sees a pattern of blue and red circles on a grid. While the examinee holds that image in working memory, he or she is shown a second grid with blue and red circles. After seeing the second grid, the examinee must add the two images together. Where there is only one blue circle across the two images, the examinee puts in a blue circle (i.e., addition), where two blue images spatially overlap, the examinee places a white circle (i.e., subtraction). The red circles must be ignored. The test is a visual analog to the WAIS-IV Arithmetic subtest and the two tests have a moderate correlation with each other ( $r = .51$ ). It measures spatial working memory and requires storage (i.e., visual sketchpad), manipulation (i.e., central executive), and ability to ignore competing stimuli (i.e., central executive).

### *Symbol Span*

Symbol Span is a visual analog to the WAIS-IV Digit Span subtest and the two tests have a moderate correlation with each other ( $r = .47$ ). The examinee sees a series of symbols that are difficult to verbalize. Subsequently, they must identify the symbols seen and identify the correct order of the symbols from left to right. The subtest measures the capacity to keep a mental image of a design in mind and the relative spatial position on the page. In [Baddeley \(2003\)](#) model, this would represent the visual sketchpad with support from the central executive (e.g., help maintain sequence and ignore competing stimuli). The test requires visual storage (e.g., increasing span) and manipulation (e.g., remembering correct sequence).

### *Auditory Memory*

The Auditory Memory Index (AMI) is composed of the Logical Memory I, Logical Memory II, Verbal Paired Associates I, and Verbal Paired Associates II scaled scores. The AMI measures the ability to listen to oral information and repeat it immediately, and then recall it again after a 20–30 min delay. The index combines measures of single trial learning and multitrial learning for verbally presented information. While a low score on the Auditory Memory Index indicates difficulties with auditory memory functioning, other cognitive problems may influence performance and need to be considered when interpreting results, including: auditory discrimination problems, language impairment, severe attention problems, poor auditory working memory, and executive functioning deficits.

Specific comparisons (e.g., simple difference or contrast scores) between AMI and WAIS-IV help the clinician determine if auditory memory deficits are a function of other cognitive limitations. The Working Memory versus Auditory Memory Index Contrast Scaled Score evaluates if auditory memory scores are lower than expected given performance on the WMI. Difficulties with working memory can have a direct impact on the examinees ability to get information in and out of long-term memory stores. These two indexes are moderately correlated ( $r = .50$ ). The

Verbal Comprehension versus Auditory Memory Index Contrast Scaled Score identifies if auditory memory scores are below expected given verbal abilities. These indexes are moderately correlation ( $r = .53$ ). The AMI requires significant receptive language skills and this contrast score can test the hypothesis that poor verbal skills are producing memory deficits or not.

### ***Logical Memory***

Logical Memory measures immediate and delayed memory for narrative stories. The examinee is read two stories and asked to recall them immediately and after a 20–30 min delay. A recognition task is also available for the delayed condition. In the Older Adult battery, one story is repeated during immediate recall. Low scores on the Logical Memory subtest indicate difficulty recalling verbal information that is conceptually organized and semantically related.

### ***Verbal Paired Associates***

Verbal Paired Associates measures immediate learning and delayed recall for word pairs. The examinee is read 10 or 14 word pairs and then given the first word of the pair and asked to recall the second. Some of the word pairs are semantically related. There are immediate and delayed cued recall conditions, a delayed recognition condition, and a free recall condition in which the examinee is asked to state all the words from the pairs without requiring the pairing. Low scores may indicate difficulties learning new associations and/or a failure to improve memory performance after multiple learning trials. The Verbal Paired Associates II score measures delayed cued recall for word associations. Low scores on Verbal Paired Associates II indicate difficulties retrieving word associations from long-term memory.

### ***Visual Memory***

The Visual Memory Index (VMI) for the Adult battery is composed of the Visual Reproduction I, Visual Reproduction II, Designs I, and Designs II scores. For the Older Adult battery, the VMI is composed only of the Visual Reproduction I and II scores. The VMI measures the ability to recall designs from memory and draw them or replicate their placement in a grid. Visual memory on WMS-IV assesses both memory for visual details and for spatial location. Like the VWMI, both spatial and visual detail content is assessed in this index, in recognition of the two distinct visual pathways that process information in the brain. While a low score on the VMI indicates difficulties with visual memory functioning, other cognitive problems may influence performance and need to be considered when interpreting results including: visual acuity deficits, visual–spatial processing impairment, severe attention problems, poor visual working memory, and impaired executive functioning.

Comparison of the VMI with the WAIS-IV can help clarify if deficits in visual memory are related to visual–spatial processing. The Perceptual Reasoning versus

Visual Memory contrast score evaluates if visual memory functioning is consistent with or different from expected performance given visual–perceptual abilities. This comparison helps identify if visual memory deficits are present beyond any visual–perceptual processing difficulties. These indexes are moderately correlated ( $r = .62$ ).

### ***Designs***

Designs measures immediate and delayed memory for spatial and visual details. The examinee is shown a series of grids with abstract visual designs in cells of the grid. After each stimulus is removed, the examinee is asked to recreate the grid with cards that contain both actual and distracter designs before being shown the next grid. Each response is scored on both correct content and correct location. Separate content and spatial location scores can calculated for both the immediate and delayed conditions. An optional recognition trial is also available.

### ***Visual Reproduction***

Visual Reproduction measures immediate and delayed memory for abstract designs. The examinee is shown a series of figures. After each stimulus is removed the examinee is asked to draw the design from memory before being presented the next design. Each design is scored for the presence of key components of the design which have elements of spatial relationship and visual details. Optional recognition and copy conditions are available to identify retrieval deficits or to identify if constructional or fine-motor issues have affected performance.

### ***Immediate Memory***

The Immediate Memory Index (IMI) measures memory for both orally and visually presented information immediately after it is presented. For the Adult battery, the immediate recall conditions of Logical Memory, Verbal Paired Associates, Designs, and Visual Reproduction are used to derive the IMI. For the Older Adult battery, the immediate recall conditions of Logical Memory, Verbal Paired Associates, and Visual Reproduction are used to derive the IMI. In the absence of significant modality specific variability in memory functioning, the IMI is the best estimate of the examinees ability to get information into long-term memory stores and immediately retrieve it. It is significantly influenced by working memory abilities as some information recalled during immediate recall tasks is present in the episodic buffer and may or may not get into long-term memory.

Comparison of IMI with WAIS-IV indexes can help refine interpretation of memory functioning in light of other contributing cognitive abilities. The Working Memory versus Immediate Memory Contrast Scaled Score tests the hypothesis that low immediate memory functioning is or is not related to poor working memory skills ( $r = .57$ ). To determine if general cognitive impairments are producing concurrent immediate memory difficulties, the General Ability versus Immediate

Memory Index Contrast Scaled Score can be evaluated ( $r = .66$ ). Low scores indicate immediate memory difficulties beyond any general cognitive impairments exhibited by the examinee.

### ***Delayed Memory***

The Delayed Memory Index (DMI) measures memory for both orally and visually presented information 20–30 min after it is presented. For the Adult battery, the delayed recall conditions of Logical Memory, Verbal Paired Associates, Designs, and Visual Reproduction are used to derive the DMI. For the Older Adult battery, the delayed recall conditions of Logical Memory, Verbal Paired Associates, and Visual Reproduction are used to derive the DMI. In the absence of modality-specific variability in memory performance the DMI is the best estimate of delayed memory functioning. The DMI is not as significantly influenced by working memory as IMI, however, general cognitive impairments can produce concomitant memory difficulties.

Comparing the WAIS-IV GAI with the WMS-IV DMI using the General Ability versus Delayed Memory Index Contrast Scaled Score can help identify if memory impairment is consistent with global cognitive impairment. Low scores on this measure indicate deficits in delayed memory functioning beyond the impact of general cognitive functioning.

### ***Brief Cognitive Status Exam***

The Brief Cognitive Status Exam (BCSE) is an optional subtest that was developed to identify examinees with significant cognitive dysfunction. Compared to the Mini Mental Status exam (Folstein, Folstein, & McHugh, 1975; Folstein, Folstein, & McHugh, 2002), the BCSE better identifies individuals with significant cognitive impairment (Hilsabeck et al., 2015). To identify atypically low performance, the BCSE uses a variety of tasks including orientation to time, mental control, planning and visual–perceptual processing, incidental recall, inhibitory control, and verbal productivity. The raw scores for each set of tasks are converted into weighted raw scores. The weighting was developed to maximize the differences between healthy controls and examinees diagnosed with dementia.

The BCSE classification levels are based on base rates of performance in the control sample: Very Low, Low, Borderline, Low Average, and Average. The classifications are not diagnostic as a number of factors contribute to performance on the subtest. If an examinee falls in the Very Low classification, they have less than a 2% chance that their score is consistent with healthy controls. Scores in the Very Low range are often obtained by examinees diagnosed with Dementia or Mild to Moderate Mental Retardation. Scores in the Low range are obtained by healthy controls approximately 2–4% of the time, such that 5% of the controls will obtain scores in the Low to Very Low range. Scores in the Low range are often obtained by examinees diagnosed with Dementia or Mild Mental Retardation. Approximately 9% of healthy controls will obtain a score in the Borderline to Very

Low range. More clinical group subjects will fall into this range as well, but the diagnostic implications are less certain. At the Borderline range and higher, there is less evidence that these scores are associated with significant cognitive impairment and the interpretation may focus on specific aspects of poor performance (e.g., poor inhibitory and mental control). The BCSE score needs to be considered in light of the clinical question and overall clinical presentation. The classification level can help practitioners provide evidence for or against significant cognitive impairment when that is an important clinical question.

## **Joint factor structure of the WAIS-IV and WMS-IV**

The conorming of WAIS-IV and WMS-IV makes it possible to explore the joint factor structure of the two batteries. Similar joint factor analytic studies have been published for the WAIS-R and WMS-R (Bowden, Carstairs, & Shores, 1999) and the WAIS-III and WMS-III (Tulsky & Price, 2003). Holdnack, Zhou, Larrabee, Millis, and Salthouse (2011) evaluated the joint factor structure of the WAIS-IV and WMS-IV. In that study, both five- and seven-factor models produced similar statistical results. The five-factor model consisted of a hierarchical general factor, Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed, and Memory. The seven-factor model did not include a hierarchical general ability factor but was comprised of Verbal Comprehension, Perceptual Reasoning, Auditory Working Memory, Visual Working Memory, Processing Speed, Auditory Memory, and Visual Memory. The five-factor model required more subtest cross-loadings to achieve best statistical fit compared to the seven-factor model. The primary difference between the two models is the emergence of modality specific factors in the seven-factor versus the five-factor model. Also, the seven-factor model produces model specification errors when a general hierarchical factor is included. The factor analysis reported by Holdnack et al. (2011) used only the delayed memory measures from the WMS-IV to avoid model specification error that can occur due to the statistical dependency between immediate and delayed memory tests. However, the same results are observed when both immediate and delayed measures are included (Holdnack et al., 2013).

Based on the results of the factor analyses, Holdnack et al. (2013) developed a new index scores combining WAIS-IV and WMS-IV working memory subtest. The new Working Memory Index is comprised of Digit Span, Arithmetic, Spatial Addition, and Symbol Span and is available for ages 16–69. As previously mentioned, a joint Quantitative Reasoning Index was created using the combined batteries. This index is comprised of Arithmetic, Figure Weights, and Spatial Addition. The index is available for individuals 16–69 years of age. Norms and clinical validation are presented in Holdnack et al. (2013).

### ***Social Perception***

Social cognition skills are required to appropriately interact with others in personal, vocational, and educational settings. Social cognition encompasses a wide variety

of skills, including affect recognition, facial memory and recognition, appropriate interpretation of affect and prosody, and theory of mind. Neurological, psychiatric, and developmental conditions (e.g., Autistic Disorder, Schizophrenia) often impact social ability or include them as diagnostic criteria. Deficits in social cognition directly impact an examinee's ability to function in most environments. In addition to specific deficits in social cognition, general cognitive deficits often produce difficulties in aspects of social cognition.

Three social cognition subtests were developed for ACS to measure aspects of social cognition, including facial affect recognition, recognition and identification of affect from prosody, ability to verbalize a speaker's intent, face recognition, and recall of names and pertinent information. Each subtest measures different clinical aspects of social cognition and can be used independently or in combination. The Social Perception subtests provide information on some basic processes involved in social cognition.

## **Advanced Clinical Solutions Social Cognition**

Social Cognition measures included in the ACS include Social Perception, Faces, and Memory for Names. Of these three measures, only Social Perception was conformed with the WAIS-IV. Social Perception measures comprehension of social communication, including facial affect recognition and naming, affect recognition from prosody and facial expressions, and affect recognition from prosody and interaction between people. Three tasks comprise Social Perception: Affect Naming, Prosody—Face Matching, and Prosody—Pair Matching. In Affect Naming, the examinee is shown photographs of faces and selects an emotion from a card to describe the affect demonstrated in the photograph. In Prosody—Face Matching, the examinee hears an audio-recorded statement and selects one face—from four choices—that matches the emotion expressed in the recording. In Prosody—Pair Matching, the examinee hears an audio-recorded statement and selects one photograph of interacting pairs of individuals—from four choices—that matches the meaning of the speaker's statement. For Prosody—Face Matching and Prosody—Pair Matching, the statement content may not match the emotion expressed. This intentional lack of matching allows for better measurement of more subtle forms of communication, such as sarcasm. Social Cognition shows good convergent and discriminant validity with other social perception measures ([Kandalaf et al., 2011](#)), and is sensitive to disorders associated with deficits in social cognition ([Kandalaf et al., 2011; Holdnack, Goldstein, & Drozdick, 2011](#)).

Performance on the Social Perception test can be affected by auditory and linguistic abilities (VCI vs SP correlation,  $r = .38$ ). Severe visual—perceptual deficits may also affect performance on this test (PRI vs SP correlation,  $r = .32$ ). In order to identify more specific deficits in social perception, the scores can be compared to WAIS-IV Indexes. The Social Perception Total Scaled score can be compared to the WAIS-IV Verbal Comprehension Index (Verbal Comprehension vs Social Perception Total Score Contrast Scaled Score). Low scores on this index indicate

deficits in social perception beyond the effects of poor language skills. Likewise, the impact of visual–perceptual deficits can be ruled out by using the Perceptual Reasoning versus Social Perception Total Scaled Score Contrast Scaled Score. Finally, if the examiner believes global cognitive issues has impacted social perception performance, General Ability versus Social Perception Total Scaled Score Contrast Scaled Score can be evaluated ( $r = .40$ ).

## **Advanced Clinical Solutions Suboptimal Effort**

Clinicians recognize the need for performance validity indicators. For a variety of reasons, an examinee may not give their best performance on testing. Threats to the validity of the test session affect the clinician's ability to accurately interpret test scores as being representative of the constructs purportedly measured by the WAIS-IV. Fluctuating or poor effort can result in lowered scores unrelated to actual deficits. Assessment of effort is increasingly included in clinical evaluations, particularly in those in which poor performance may benefit the examinee. Assessing effort can be difficult because many factors can influence effort, including medications, fatigue, motivation, and depression or other psychological disorders. Multiple sources of information are required to determine and validate if an individual is giving suboptimal effort. The ACS includes a new measure of effort, Word Choice, and normative and comparative information on embedded measures of effort in the WAIS-IV (e.g., Reliable Digit Span) and the WMS-IV (e.g., Logical Memory Recognition, Verbal Paired Associates Recognition, and Visual Reproduction Recognition).

### ***Word Choice***

The Word Choice subtest is a forced choice memory test. Examinees are read and shown a list of 50 words and then asked to select the word they heard from two choices. Typically developing individuals obtain nearly perfect scores. Information on chance performance by an examinee is provided as well as performance data from multiple clinical samples, including simulator (i.e., individuals asked to feign poor performance) and no stimulus groups (i.e., individuals who were never shown the words). An examinee's performance can be compared to these data to determine if their performance is worse than performance observed in true clinical populations.

## **Refining interpretation of the WAIS-IV**

The ACS contains several procedures that enables the clinician to test specific hypotheses about the WAIS-IV scores obtained by a specific examinee. These procedures help clinicians identify a change in cognitive functioning. Specifically, the

Demographic Normative Adjustments and the Test of Premorbid Functioning (TOPF) indicate if a current score is lower than expected given the examinee's background. The Reliable Change Index is used to compare performance to determine if a reevaluation shows a significant decline from the previous evaluation.

### ***Demographic referenced norms***

Clinicians frequently evaluate patients from diverse socioeconomic and cultural backgrounds. When evaluating patients from minority or economically disadvantaged environments, the clinician needs to consider the extent to which cultural, educational, financial, and other environmental factors, such as access to healthcare, impact performance on cognitive tests. An accurate diagnosis requires the clinician to discern the degree to which an achieved low score represents the effects of a disease process rather than the impact of cultural, educational, and economic factors. Similarly, the extent to which a high level of educational achievement and financial status impact test performance also needs to be considered, particularly in the presence of possible cognitive loss or decline where average scores may indicate a loss of function. In these evaluations, the clinician attempts to determine if an achieved score represents an *expected* level of performance for that *individual* based on their background characteristics.

Norms provided a standardized means to compare an individual's performance to a group of peers, typically same-age peers. When the purpose of the evaluation is to evaluate the examinee in relation to the general population, standard age-referenced norms are the most appropriate and psychometrically robust method of comparison. However, some clinical questions require the examiner to determine whether an examinee has experienced a change in functioning or to compare the examinee to individuals with similar background characteristics (e.g., education level). For these types of evaluations, demographically referenced norms may be appropriate. Demographic referenced norms compare an individual's performance to individuals with similar background characteristics (e.g., education level). The ACS applies demographic referencing (i.e., education only, full demographics) to the standard age-referenced WAIS-IV and WMS-IV subtest and composite scores.

The user of these norms should be aware that the demographic-referenced scores are the normed residual of the predicted versus actual performance. This model yields very predictable results (e.g., higher scores are returned for lower education levels). The adjusted scores represent the distance from the mean of individuals with a similar background (e.g., white, female, 18 years of education). Scores well below the mean indicate lower than expected performance compared to similar others and in some cases may signal decline in functioning. There are many caveats to using these scores and these are detailed in Holdnack et al. (2013).

### ***Test of Premorbid Functioning***

Clinicians are frequently asked to determine if an examinee's current functioning is lower than their functioning prior to an injury or onset of neurological disease

process. Current performance does not provide information on how current performance related to prior ability. For example, low performance on a current assessment does not indicate a decline because the examinee may have had low ability prior to the injury. For most individuals, testing was not completed prior to these injuries and the examiner is required to estimate the examinee's prior functioning. Historically, an individual's background characteristics (e.g., education, occupation) were used to estimate premorbid ability. However, this approach is prone to error due to estimation error and bias. Modern approaches to estimating premorbid functioning involve statistical approaches to providing estimates of ability based on current performance on tasks that are known to be minimally affected by cognitive change, at least in individuals with mild to moderate impairment, or on demographic comparisons.

The TOPF was developed for the ACS to estimate an individual's premorbid WAIS-IV and WMS-IV scores. The TOPF is a reading task in which the examinee reads aloud words that have irregular grapheme-to-phoneme translation. It is a revision of the Wechsler Test of Adult Reading (WTAR: [Pearson, 2001](#)) and shares methodology and some content with the National Adult Reading Test (NART: [Nelson, 1982](#)), the American Version of the National Adult Reading Test (AMNART: [Grober & Sliwinski, 1991](#)), and the North American Adult Reading Test (NAART: [Blair & Spreen, 1989](#)). This type of reading task is less affected by cognitive disease processes and injury than other measures of intellectual and memory functioning.

The TOPF can be used alone or in conjunction with demographic characteristics to provide estimates of premorbid ability. Complex multivariate regression equations are used to derive the expected level of performance. The examiner compares the difference between expected performance and actual performance to determine if the scores are significantly different than expected and statistically rare. Significant and rare differences may signal a decline in cognitive functions. There are many caveats to using this approach which are detailed in [Holdnack et al. \(2013\)](#).

An alternative prediction model is to use subtests from within the WAIS-IV (e.g., Vocabulary and Matrix Reasoning) that are relatively resistant to the effects of brain injury or cognitive decline ([Holdnack et al., 2013](#)). The Oklahoma Premorbid Intelligence Estimate (OPIE-3: [Schoenberg, Duff, Scott, & Adams, 2003](#); [Schoenberg, Scott, Duff, & Adams, 2002](#)) was originally developed for the WAIS-III and is available for the WAIS-IV ([Holdnack et al., 2013](#)). This model combines demographic data with performance on Vocabulary, Matrix Reasoning, or the combination of these measures to estimate expected level of performance. Like the TOPF, the OPIE-IV uses regression equations to predict expected levels of performance. The difference between predicted and actual performance is evaluated for statistical significance and statistical rarity. Cognitive decline may be indicated in cases where actual versus predicted scores are statistically significant and rare in normal controls. There are many caveats to using this approach which are detailed in [Holdnack et al. \(2013\)](#).

### ***Serial assessment with WAIS-IV and WMS-IV***

In clinical and research settings, individuals are routinely reevaluated with the same instruments. Typically this is done to determine if a change in ability has occurred as part of an ongoing disease process (e.g., dementia) or following treatment or intervention. However, a number of statistical and personal effects need to be accounted for before a change in performance can be attributed to a change in actual ability. Scores on reevaluation need to be considered within this context in order to avoid incorrect interpretation of changes in scores on retest. The ACS serial assessment scores are adjusted for the statistical and personal effects known to influence retest scores in normally developing individuals. Use of these scores greatly increases the likelihood of correctly interpreting the relation between changes in scores and actual changes in ability.

There are a number of factors that influence changes in test score overtime. These include psychometric factors such as the internal reliability of the test, the retest reliability, and regression to the mean effects. Additionally, the nature of the test itself can influence score differences overtime, in particular tests that rely on novelty, learning, or procedural learning can yield significantly higher scores on retesting. This practice or exposure effect has direct implications for comparing tests taken at two different times. Furthermore, characteristics of the individual can play a role in the degree to which practice effects are observed. Factors such as age, sex, ability level, and education level must be considered when interpreting performance change. The ACS uses multivariate regression equations that control for individual characteristics and practice effects to determine an expected level of performance at time 2 given performance at time 1. The examiner compares the current performance to the expected performance. If the difference is both statistically significant and statistically rare in healthy controls, then the performance change may indicate a loss in functioning over the period of the time 1 to time 2 assessment. There are a number of caveats to applying this methodology, which are detailed in [Holdnack et al. \(2013\)](#).

### ***Multivariate base rates***

Research on multivariate base rates has illustrated and emphasized that when healthy people complete a battery of tests, a substantial minority will obtain one or more low scores ([Brooks, 2010, 2011; Brooks, Holdnack, & Iverson, 2011; Brooks, Iverson, Holdnack, & Feldman, 2008; Iverson, Brooks, & Holdnack, 2008; Schretlen, Testa, Winicki, Pearson, & Gordon, 2008](#)). This extensive line of research has provided clinicians with information for interpreting test performance in adults and older adults on several large batteries of conormed neuropsychological tests, the Wechsler Adult Intelligence Scale—third edition/Wechsler Memory Scale—third edition conormed battery (WAIS-III/WMS-III; [Iverson et al., 2008](#)), and the Wechsler Adult Intelligence Scale—fourth edition/Wechsler Memory Scale—fourth edition conormed battery (WAIS-IV/WMS-IV: [Brooks et al., 2011](#)).

The multivariate base rate effect is illustrated by identifying the number of healthy individuals that obtain a score at or below a specific cutoff. When considering a subtest scaled score of 6, approximately 9% of individuals will have that score or lower when considering that test score in isolation. However, if one considers the test as part of a larger battery including all 20 subtests from the WAIS-IV/WMS-IV, 61.1% of healthy controls will obtain at least 1 or more scaled scores of 6. A failure to account for the phenomenon can result in overinterpreting one or more low scores as indicative of cognitive impairment or atypical cognitive functioning.

It has been demonstrated that using multivariate base rate tables can improve sensitivity and specificity in identifying atypical cognitive performance (Brooks et al., 2011; Brooks, Iverson, Feldman, & Holdnack, 2009; Holdnack et al., 2013). When base rates are further refined by premorbid ability to demographic characteristics, the sensitivity and specificity are further improved. Additionally, base rates by domain are now available to aid in identifying specific domains of cognitive impairment (Holdnack et al., 2013).

### **Cognitive variability**

Clinicians frequently compare performance between tests and subtests in order to determine if a patient exhibits significant strengths or weaknesses in their cognitive functioning as part of psychological, psychoeducational, and neuropsychological evaluations. The practice of interpreting test scatter or the relative differences among and between various scores associated with psychometric tests is steeped in a rich historical and clinical tradition. The interpretation of test scatter has historically been used as a tool to help in determining the presence of cognitive abnormality and to aid in developing hypotheses about an individual's pattern of cognitive strengths and weaknesses (Kaufman & Lichtenberger, 2006; Lezak et al., 2004; Sattler & Ryan, 2009). Statistical investigations into determining normal and abnormal levels of test scatter on certain Wechsler tests date back more than 40 years (Kaufman, 1976). However, these early investigations were related primarily to the cognitive performance of children on the WISC-R. Some of the earliest scatter norms for the WAIS-R were published approximately 10 years later (Matarazzo, Daniel, Prifitera, & Herman, 1988; Matarazzo & Herman, 1985).

A high degree of variability among cognitive skills, or specific types of variability, are thought to be associated with specific clinical populations or are considered a general indicator of abnormal brain functioning. Given the large quantity of pairwise comparisons that could be made in a battery of tests, the probability of a type 1 error is inflated relative to the specific significance level used to identify differences between two test scores (i.e.,  $p$  value of .05 applied only to a single comparison not to multiple comparisons). When considering the highest versus lowest scaled score on the 10 WAIS-IV subtests, the average difference in the standardization sample is 6.5 scaled score points, a value more than 2 standard deviations from the mean. Approximately 16% of the standardization sample have a highest versus lowest score of 3 or more standard deviations (Holdnack et al., 2013). The

probability of observing a significant difference is very high and is directly related to the highest obtained score. High scoring individuals will yield more significant differences than low scoring individuals. As a result, individuals with cognitive problems (e.g., most clinical examinees) will show fewer significant differences and less overall cognitive variability. It is possible to identify significant cognitive variability when controlling for the effects of high versus low scores (Holdnack et al., 2013). Normative tables have been developed to help clinicians identify unusual levels of cognitive variability using the WAIS-IV, WMS-IV, and combined WAIS-IV/WMS-IV (Holdnack et al., 2013).

## Future directions

It is not possible to know what the exact direction that revisions to the WAIS-IV will take. However, it is possible based on current knowledge to speculate as to the potential revisions of the WAIS-IV to the WAIS-V. Based on the publication of the WISC-V and factor analytic research, there is a very high probability that the WAIS-V will move to a five-factor model that includes: Verbal Comprehension, Fluid Reasoning, Visual–Spatial Reasoning, Working Memory, and Processing Speed. There is a high probability that the WAIS-V will be conformed with the revision of the WMS-IV (e.g., WMS-V). Additional tests such as Social Perception, TOPF, Word Choice, and measures of executive functioning will be incorporated into the revision cycle. It is expected that there will be an emphasis on streamlining subtests, reducing nonconstruct related variance, and reducing administration time.

The biggest game-changer in the development of future Wechsler Scales is the Q-interactive platform. The digital administration allows for novel tests to be developed and refined that could never be accurately administered in paper and pencil format. It should be anticipated that digitally native measures will be created for the new edition. This does not necessarily suggest that a paper and pencil edition would not be developed, rather, it would be expected that the revision will leverage the power of the new administration platform in ways that have only begun with the publication of the WISC-V.

## References

- Adair, J. C., Na, D. L., Schwartz, R. L., & Heilman, K. M. (1998). Analysis of primary and secondary influences on spatial neglect. *Brain and Cognition*, 37, 351–367.
- Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews/Neuroscience*, 4, 829–839.
- Bate, A. J., Mathias, J. L., & Crawford, J. R. (2001). Performance on the Test of Everyday Attention and standard tests of attention following severe traumatic brain injury. *The Clinical Neuropsychologist*, 15, 405–422.

- Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of Wechsler Adult Intelligence Scale-fourth edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment, 22*, 121–130.
- Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Anée Psychologique, 11*, 191–244.
- Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the national adult reading test. *The Clinical Neuropsychologist, 3*, 129–136.
- Bodin, D., Pardini, D. A., Burns, T. G., & Stevens, A. B. (2009). Higher order factor structure of the WISC-IV in a clinical neuropsychological sample. *Child Neuropsychology, 15*, 417–424.
- Bowden, S. C., Carstairs, J. R., & Shores, E. A. (1999). Confirmatory factor analysis of combined Wechsler Adult Intelligence Scale-Revised and Wechsler Memory Scale-Revised scores in a healthy community sample. *Psychological Assessment, 11*, 339–344.
- Brooks, B. L. (2011). A study of low scores in Canadian children and adolescents on the Wechsler Intelligence Scale for children-fourth edition (WISC-IV). *Child Neuropsychology, 17*, 281–289.
- Brooks, B. L., Holdnack, J. A., & Iverson, G. L. (2011). Advanced clinical interpretation of the WAIS-IV and WMS-IV: Prevalence of low scores varies by level of intelligence and years of education. *Assessment, 18*, 156–167.
- Brooks, B. L., Iverson, G. L., Feldman, H. H., & Holdnack, J. A. (2009). Minimizing misdiagnosis: Psychometric criteria for possible or probable memory impairment. *Dementia and Geriatric Cognitive Disorders, 27*, 439–450.
- Brooks, B. L., Iverson, G. L., Holdnack, J. A., & Feldman, H. H. (2008). Potential for misclassification of mild cognitive impairment: a study of memory scores on the Wechsler Memory Scale-III in healthy older adults. *Journal of the International Neuropsychological Society, 14*, 463–478.
- Chen, H., Keith, T., Chen, Y., & Chang, B. (2009). What does the WISC-IV measure? Validation of the scoring and CHC-based interpretative approaches. *Journal of Research in Education Sciences, 54*, 85–108.
- Chen, H., Keith, T., Weiss, L., Zhu, J., & Li, Y. (2010). Testing for multigroup invariance of second-order WISC-IV structure across China, Hong Kong, Macau, and Taiwan. *Personality and Individual Differences, 49*, 677–682.
- Chen, H., & Zhu, J. (2012). Measurement invariance of WISC-IV across normative and clinical samples. *Personality and Individual Differences, 52*, 161–166.
- Daniel, M. H. (2007). “Scatter” and the construct validity of FSIQ: Comment on Fiorello et al. (2007). *Applied Neuropsychology, 14*, 291–295.
- Daniel, M. (2012). Equivalence of Q-interactive—Administered Cognitive Tasks: WAIS-IV. [http://www.helloq.com/content/dam/ped/ani/us/helloq/media/QinteractiveTechnical%20Report%201\\_WAIS-IV.pdf](http://www.helloq.com/content/dam/ped/ani/us/helloq/media/QinteractiveTechnical%20Report%201_WAIS-IV.pdf).
- Fiorello, C. A., Hale, J. B., Holdnack, J. A., Kavanagh, J. A., Terrell, J., & Long, L. (2007). Interpreting intelligence test results for children with disabilities: Is global intelligence relevant? *Neuropsychology, 14*, 2–12.
- 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.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (2002). *Mini-Mental State Examination*. Lutz, FL: Psychological Assessment Resources.
- Galton, F. (1869). *Heredity genius: An inquiry into its laws and consequences*. London: MacMillan.

- Geldmacher, D. S., Fritsch, T., & Riedel, T. M. (2000). Effects of stimulus properties and age on random array letter cancellation tasks. *Aging, Neuropsychology, and Cognition*, 7, 194–204.
- Goddard, H. H. (1910). A measuring scale of intelligence. *Training School*, 6, 146–155.
- Gottfredson, L. S. (1997). Why *g* matters: the complexity of everyday life. *Intelligence*, 24, 79–132.
- Gottfredson, L. S. (1998). The general intelligence factor. *Scientific American Presents*, 9, 24–29.
- Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology*, 13, 933–949.
- Hilsabeck, R. C., Holdnack, J. A., Cullum, M. C., Drozdick, L. W., Edelstein, B., Fiske, A., ... Wahlstrom, D. (2015). The Brief Cognitive Status Examination (BCSE): Comparing Diagnostic Utility and Equating Scores to the Mini Mental State Examination (MMSE). *Archives of Clinical Neuropsychology*, 30, 458–467.
- Holdnack, J. A., Drozdick, L. W., Weiss, L. G., & Iverson, G. L. (2013). *WAIS-IV/WMS-IV/ACS: Advanced clinical interpretation*. San Diego, CA: Academic Press.
- Holdnack, J., Goldstein, G., & Drozdick, L. (2011). Social Perception and WAIS-IV Performance in adolescents and adults diagnosed with Asperger's Syndrome and Autism. *Assessment*, 18, 192–200.
- Holdnack, J. A., Zhou, X., Larrabee, G. J., Millis, S. R., & Salthouse, T. A. (2011). Confirmatory Factor Analysis of the WAIS-IV/WMS-IV. *Assessment*, 18, 178–191.
- Iverson, G. L., Brooks, B. L., & Holdnack, J. A. (2008). Misdiagnosis of cognitive impairment in forensic neuropsychology. In R. L. Heilbronner (Ed.), *Neuropsychology in the courtroom: Expert analysis of reports and testimony* (pp. 243–266). New York: Guilford Press.
- Jackson, D. N. (1998). *Multidimensional aptitude battery-II*. Port Huron, MI: Sigma Assessment Systems.
- Kandalaf, M. R., Didehbani, N., Cullum, C. M., Krawczyk, D. C., Toon, T. R., Tamminga, C. A., & Chapman, S. B. (2011). The Wechsler ACS social perception subtest: A preliminary comparison with other measures of social cognition. *Journal of Psychoeducational Assessment*, 20, 455–465.
- Kaufman, A. S. (1976). Verbal-performance IQ discrepancies on the WISC-R. *Journal of Consulting and Clinical Psychology*, 44, 739–744.
- Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). Hoboken, NJ: Wiley.
- Keith, T. Z., Fine, J. G., Taub, G., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children-fourth edition: What does it measure? *School Psychology Review*, 35, 108–127.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, 86, 148–161.
- Lecerf, T., Rossier, J., Favez, N., Reverte, I., & Coleaux, L. (2010). The four versus alternative six factor structure of the French WISC-IV: Comparisons using confirmatory factor analyses. *Swiss Journal of Psychology*, 69, 221–232.
- Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment* (4th ed.). New York: Oxford University Press.

- Lichtenberger, E., & Kaufman, A. S. (2009). *Essentials of WAIS-IV assessment*. New York: Wiley.
- Lindemann, J. E., & Matarazzo, J. D. (1990). Assessment of adult intelligence. In G. Goldstein, & M. Hersen (Eds.), *Handbook of psychological assessment* (2<sup>nd</sup> Ed.). Elmsford, NY: Pergamon Press.
- Lohman, D. F. (2011). *Cognitive abilities test, form 7 (CogAT7)*. Rolling Meadows, IL: Riverside Publishing.
- Matarazzo, J. D., Daniel, M. H., Prifitera, A., & Herman, D. O. (1988). Inter-subtest scatter in the WAIS-R standardization sample. *Journal of Clinical Psychology*, 44, 940–950.
- Matarazzo, J.D. & Herman, D.O. (1985). Clinical uses of the WAIS-R: Base rates of differences between VIQ and PIQ in the WAIS-R standardization sample. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications* (pp. 899–932).
- Na, D. L., Adair, J. C., Kang, Y., Chung, C. S., Lee, K. H., & Heilman, K. M. (1999). Motor Perseverative behavior on a line cancellation task. *Neurology*, 52, 1569–1576.
- Nelson, H. E. (1982). *National adult reading test*. Windsor: NFER-Nelson.
- Pearson. (2001). *Wechsler test of adult reading*. San Antonio, TX: Author.
- Pearson. (2009). *Advanced Clinical Solutions for use with WAIS-IV and WMS-IV*. San Antonio, TX: Author.
- Prifitera, A., Saklofske, D. H., & Weiss, L. G. (2005). *WISC-IV clinical use and interpretation: Scientist-practitioner perspectives*. San Diego: Academic Press.
- Prifitera, A., Saklofske, D. H., & Weiss, L. G. (2008). *WISC-IV Clinical Assessment and Intervention* (2nd ed.). San Diego: Academic Press.
- Prifitera, A., Weiss, L. G., & Saklofske, D. H. (1998). The WISC-III in context. In A. Prifitera, & D. H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 1–38). San Diego: Academic Press.
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven's Progressive Matrices and Vocabulary Scales*. Oxford, UK: Oxford Psychologists Press.
- Reynolds, C. R., & Kamphaus, R. W. (2015). *Reynolds Intellectual Assessment Scales* (2nd ed.). Lutz, FL: Psychological Assessment Resources.
- Sattler, J. M., & Ryan, J. J. (2009). *Assessment with the WAIS-IV*. La Mesa, CA: Jerome M. Sattler Publisher.
- Schoenberg, M. R., Duff, K., Scott, J. G., & Adams, R. L. (2003). An evaluation of the clinical utility of the OPIE-3 as an estimate of premorbid WAIS-III FSIQ. *The Clinical Neuropsychologist*, 17, 308–321.
- Schoenberg, M. R., Scott, J. G., Duff, K., & Adams, R. L. (2002). Estimation of WAIS-III intelligence from combined performance and demographic variables: Development of the OPIE-3. *The Clinical Neuropsychologist*, 16, 426–438.
- Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock–Johnson IV tests of cognitive abilities examiner's manual, standard and extended batteries*. Itasca: Riverside.
- Schretlen, D. J., Testa, S. M., Winicki, J. M., Pearlson, G. D., & Gordon, B. (2008). Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. *Journal of the International Neuropsychological Society*, 14, 436–445.
- Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.
- Tulsky, D. S., & Price, L. R. (2003). The joint WAIS-III and WMS-III factor structure: Development and cross-validation of a six-factor model of cognitive functioning. *Psychological Assessment*, 15, 149–162.

- Watkins, M. W., Glutting, J. J., & Lei, P. W. (2007). Validity of the full-scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology, 14*, 13–20.
- Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore, MD: Williams & Witkins.
- Wechsler, D. (1997a). *Wechsler adult intelligence scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997b). *Wechsler memory scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). San Antonio, TX: Pearson, Inc.
- Wechsler, D. (2008). *Wechsler adult intelligence scale* (4th ed.). San Antonio, TX: Pearson, Inc.
- Wechsler, D. (2009). *Wechsler memory scale* (4th ed.). San Antonio, TX: Pearson, Inc.
- Wechsler, D. (2013). *Wechsler intelligence scale for children* (5th ed.). San Antonio, TX: Pearson, Inc.
- Weiss, L. G., Keith, T. Z., Chen, H., & Zhu, J. (2013). WAIS-IV: Clinical validation of the four- and five-factor interpretive approaches. *Journal of Psychoeducational Assessment, 31*, 94–113.
- Weiss, L. G., Saklofske, D. H., Coalson, D., & Raiford, S. E. (2010). Theoretical, empirical, and clinical foundations of the WAIS-IV index scores. In L. G. Weiss, D. H. Saklofske, D. Coalson, & S. E. Raiford (Eds.), *WAIS-IV: Clinical use and interpretation: Scientist-practitioner perspectives*. San Diego: Academic Press.
- Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (2006). *WISC-IV: Advanced clinical interpretation*. San Diego: Academic Press.
- Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests*. New York: Henry Holt and Company.

## Further reading

- Coalson, D. L., Raiford, S. E., Saklofske, D. H., & Weiss, L. G. (2010). WAIS-IV: Advanced assessment of intelligence. In L. G. Weiss, D. H. Saklofske, D. Coalson, & S. E. Raiford (Eds.), *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives*. San Diego: Elsevier Science.

## **Part IV**

# **Achievement and Interest**

# Aptitude and achievement testing

5

Lynda J. Katz<sup>1</sup> and Franklin C. Brown<sup>2</sup>

<sup>1</sup>Private Practice, Durham, NC, United States, <sup>2</sup>Yale University, Department of Neurology, New Haven, CT, United States

## Aptitude and achievement testing

While the uses of both achievement and aptitude testing have undergone a series of, at times, dramatic changes in the sociopolitical educational environment since the publication of the *Handbook of Psychological Assessment* (third edition) in 2000 (Katz & Slomka, 1990), the basic nature of the assessment instruments used in clinical practice has not. Therefore it is our intent to briefly review the definitions of both achievement and aptitude testing from a historical perspective and then focus in greater detail on the uses of these measures in the past 10–15 years and their current use. While the basic nature of these tests has not changed, other than some useful updates, clinicians have had to more carefully consider the applications and standards of practice in their use. This is particularly important when evaluating those with diverse backgrounds who have been referred for or have sought out psychological and/or neuropsychological assessments.

In her 1984 paper, Anastasi described aptitude and achievement tests as “The Curious Case of the Indestructible Strawperson” in which she points out the extensive overlap and lack of distinction between these two types of tests, and then goes on to explicate the concept that “both aptitude and achievement tests can be best characterized as tests of developed ability” (p. 134). She then draws the distinction between the two in terms of the experiential pool used in formulating the various test items: “tests of developed ability differ in the degree of precision versus vagueness with which the relevant domain of antecedent experience is defined” (p. 136). Finally, while traditionally aptitude tests were designed and used to predict future performance following a specific learning experience, achievement tests were designed and used primarily to assess the current status of the individual or groups of individuals (Anastasi, 1984). However, in essence both measures assess current status while aptitude tests focus more specifically on the predictive nature of test data gleaned from their administration. This distinction holds true today in term of the uses for which both kinds of assessment measures are employed.

Therefore we first will address the current uses of aptitude testing in light of the distinctions which Anastasi has drawn. Then, in more detail, we will address the use of achievement testing in clinical practice as this is judged to be an important

component of a psychoeducational and neuropsychological assessment in the evaluation of childhood learning disorders (Silver et al., 2006) and, we posit, in the evaluation of adolescents and adults as well.

## Aptitude testing

Traditionally, the application of aptitude testing has depended to some extent on whether an aptitude has been judged to be something that is stable over time or whether it is modifiable (Stemler & Sternberg, 2013). As a compromise position, Silzer and Church (2009) proposed that certain components of aptitude represent foundational dimensions and other components represent growth dimensions. In other words, foundational dimensions are relatively stable traits such as strategic thinking and interpersonal skills, whereas such things as openness to feedback, risk-taking, and achievement orientation can be developed and expanded (Silzer & Church, 2009). Thus the assessment strategies may differ to some degree, depending on the conceptual orientation of those conducting assessments of aptitude. This is particularly the case in certain areas such as college admissions testing and employment testing, including for the military, where there has been an evolution in the way in which aptitude testing has been conceptualized and applied over the years since its inception.

## College and professional schools

The original purpose of college admissions testing was to identify those students who would be best able to profit from postsecondary education using a set of test scores derived from an intelligence-like test. In 1925 the College Board made a standardized Scholastic Aptitude Test (SAT) available to colleges, essentially based on a testing of general reasoning ability, drawing heavily on the intelligence testing that had arisen in the United States by that time. The SAT was based on a narrow range of cognitive abilities (verbal and quantitative reasoning) theorized to be domain-general traits residing within a person (Stemler & Sternberg, 2013). After some years of debate as to whether the test measured aptitude, ability, achievement, or a combination of these factors, the title of the test was changed to the Scholastic Assessment Test and then, finally, simply the SAT, with the prior names abandoned. The combined SAT score correlates with first-year college grade point average (GPA) at 0.53. Controlling for high school GPA provides an additional increment of 0.08 to the predictive validity of the test (Halpern & Butler, 2013). This seems like a very small value for incremental validity, but nonetheless some argue that the small increase in explained variance has a meaningful effect when predicting the percentage of students who succeed in college (Bridgman, Pollack, & Burton, 2004).

The other most commonly used test in the college admission process, the American College Testing (ACT, 2007), was originally created as an extension of

the Iowa testing programs to serve the postsecondary community and as such was considered to be a measure of achievement. The original measure consisted of a battery of four multiple-choice tests in English usage, mathematics usage, social sciences reading, and natural sciences reading (Peterson, 1983). Since 1989 however, the separate reading measures were combined resulting in a single reading test and a new science test. After significant studies of its content validity, it has been adopted by numerous states as part of the testing requirements established by the “No Child Left Behind Act” (2001) legislation (Lissitz & Samuelsen, 2007). Today’s ACT is a nearly 3-hour examination that reports scores in all four areas as well as a combined score (ACT, 2007).

In addition to validity concerns that have arisen over the years with the use of the SAT and the ACT, the matter of test fairness as both “a statistical concept and an issue of values” (Halpern & Butler, 2013) has been hotly debated with respect to college entrance examinations. Ethnic and racial group differences on the SAT have been reported (Patterson, Mattern, & Kobrin, 2009), with as much as a one standard deviation difference between African Americans and Caucasians, with smaller differences in scores found for every other ethnicity examined (Kobrin, Carmase, & Milowski, 2002). In keeping with the controversy, Helms (2006) argued that while validity evidence is necessary for test fairness, it is not sufficient. She continued that racial group is used as a proxy for socialization experiences that affect the way individuals react to the testing situation, and any test with group differences in average scores is unfair (Helms, 2006). In addition, gender differences have been identified on the SAT with women outperforming men on some tests of verbal ability, especially writing, but underperforming on tests of visuospatial skills and quantitative ability (Halpern et al., 2007).

And while concerns about the validity and fairness of college admission assessments have led a number of colleges and universities to exclude or deemphasize their importance, these standardized assessments continue to be used by the vast majority of higher education institutions including graduate and professional programs (GRE, MCAT, GMAT, LSAT, etc.). According to Halpern and Butler (2013), there is no evidence, in general, regarding group differences in the predictive validity of the standardized graduate school entrance examinations based on gender or ethnicity, but if differences were found, they tended to favor ethnic minorities (Halpern & Butler, 2013).

In one interesting study (Collin, Violato, & Hecker, 2009) regarding the predictive validity of the Medical College Admission Test (MCAT), data were gathered on over 800,000 participants from 1991 to 2000 including scores on the MCAT, undergraduate grade point averages (UGPA), and results from steps 1, 2, and 3 of the United States Medical Licensure Examination (USMLE). These three measures were able to identify three theoretical constructs, aptitude for medicine (the MCAT is assumed to assess this construct), general achievement (UGPA), and competence in medicine resulting in a comparative fit index = 0.932 based on a sample of 20,714 individuals. While there were some differences in MCAT scores between men and women, with men scoring higher on the hard sciences subtests, females outperformed males on steps 2 and 3 of the USMLE (given at the end of medical

school and after 1 year of residency). Age at admission, as a variable, was inversely related to undergraduate GPA, both in the sciences and nonscience courses, but not to MCAT subtest scores or USMLE step 1 and 2 scores. The authors of the study concluded:

*What characteristics of candidates predict the success or performance of a medical student and resident? We can say that the present study answered part of that question—that is cognitive-achievement related variables are an important aspect of success throughout both undergraduate and postgraduate medical education.*

*Collin et al. (2009), p. 365.*

Finally, it is important to note for the clinician working with students with disabilities that performance on these measures can dictate whether or not an individual may pursue a vocational or professional career aspiration. And so the issue of whether or not reasonable accommodations are justified under the [Americans With Disabilities Act \(1990\)](#) becomes a major factor in the assessment process, regardless of the theoretical underpinning of the instrument used.

## **Vocational preparation and employment**

Vocational preparation, employment, and personnel selection are other arenas in which the measurement of aptitude has a long history. Understanding the underlying structure of the various measures of aptitude can be found in the work of [Carroll \(1993\)](#) who defined a hierarchical model of intelligence. This consisted of three levels. General intelligence is at the top of the hierarchy. The second level consists of fluid and crystallized intelligence, memory, retrieval, visual and auditory perception, cognitive speed, and processing speed. Individual tests of aptitude and achievement fall at the third level ([Carroll, 1993](#)).

Among these measures of aptitude, one finds significant commonality in construct formation. Specifically, the measurement of domain-specific abilities includes verbal reasoning; word knowledge and language usage skills; arithmetic reasoning and computation skills; spatial relationships; form perception; manual and/or finger dexterity; and motor coordination. Examples of attempts to assess those underlying abilities or aptitudes that have been demonstrated to predict success in specific occupational clusters are found in the work of researchers such as [Schmidt and Hunter \(1998\)](#) and [Webb, Lubinski, and Benbow \(2002\)](#). As a testimony to the utility of general abilities, [Schmidt and Hunter \(1998\)](#) conducted a meta-analysis to examine the validity of measures to predict job performance in psychology over an 85-year period. They found that the combination of general mental ability (e.g., intelligence), work samples, and a structured interview resulted in the greatest ability to predict job performance. Tests of conscientiousness, reference checks, and

peer ratings were correlated with job performance, but did not add significant yield beyond these three methods, whereas interest inventories, years of education or job experience, and hand writing had no predictive ability for job performance. [Webb et al. \(2002\)](#) found that specific math skills demonstrated during early teen years predicted later job satisfaction and performance in careers involving math and science. Interestingly, these early skills predicted job choice more than actual college majors (e.g., some majored outside of math and sciences, but later returned to those areas within their careers). Thus overall abilities predict success, regardless of the field; and more specific academic abilities are a strong predictor of career performance and satisfaction.

With respect to vocational preparation and personnel selection, several major measures of aptitude continue to be utilized by secondary schools, rehabilitation agencies, and human resource management personnel. These include the COPSSystem, which includes the Career Occupational Preference System Interest Inventory (COPS), Career Ability Placement Survey (CAPS), and Career Orientation Placement and Evaluation Survey (COPES) ([Knapp, Knapp, & Knapp-Lee, 1977](#)). Other measures include the Differential Aptitude Test (DAT) ([Bennett, Seashore, & Wesman, 1947](#)) and the General Aptitude Test Battery (GATB), which can be purchased from the US Government Printing Office. The CAPS measures abilities related to various occupational clusters. It is a user-friendly instrument which can be self-scored or can be administered and scored via the internet. Comparative validity data between the CAPS and the GATB were established and reported in a 1978 study ([Knapp, Knapp, Strnad, & Michael, 1978](#)). Their comparability was also demonstrated with a sample of psychiatric patient referrals in a vocational rehabilitation program ([Katz, Beers, Geckle, & Goldstein, 1989](#)). Concurrent validity studies were also conducted with the DAT and ranged from 0.65 to 0.81 between conceptually similar tests ([Knapp et al., 1977](#)). The DAT ([Bennett et al., 1947](#)) has been recently renamed the Differential Aptitude Tests for Personnel and Career Assessment, which is available through the Psychological Corporation.

In addition, there are a number of measures intended to assess specific ability in the area of mechanical aptitude. Among them are the Bennett Mechanical Comprehension Test (BMCT) ([Bennett, 1940, 1994, 2008](#)). The manual for the most recent edition can be found online and published through [NCS Pearson, Inc. \(2008\)](#). Another test, the Wiesen Test of Mechanical Aptitude (WTMA; [Wiesen, 1977](#)), is available through the company he founded, APR Testing Services. Both are used to measure an individual's knowledge of mechanical and physical concepts. These tests have been found to be useful in predicting success in occupations requiring mechanical aptitude, spatial visualization, the application of physics, and deduction of "how things work" (<https://www.testprep-online.com.bmct.aspx>). In contrast, a more recent online measure is the Ramsay Mechanical Aptitude Test-Form Mat-4 ([Ramsay Corporation, 2004](#)), designed to evaluate a person's ability to learn production and maintenance job activities but not to measure specific knowledge and skills. The intent is to predict success in apprenticeship or trainee programs.

## Military testing for identification and classification

The military, in particular, has continued to conduct research based on the concept of aptitude as a domain-specific construct that requires a match between the demands of a particular occupation and aptitudes of a perspective candidate for that position. They have moved from a single measure (the Army Alpha and Beta developed in 1917) to a more differentiated multiple aptitude measure, the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB was introduced in the 1970s and replaced measures developed separately by the Army, Navy, and Air Force (Rumsey & Arabian, 2014). Although two types of noncognitive measures have endured over the years, namely, personality/temperament (terms used interchangeably) and vocational interest testing, the focus on cognitive measures more clearly reflects the notion of domain-specific constructs.

The ASVAB is a multiple-choice test used to determine qualification for enlistment in the US Armed Forces (US Department of Defense, 1984). It currently contains nine sections, with each test varying in length—the longest being 36 minutes for the Arithmetic Reasoning test—for a total of 3 hours. The ASVAB is administered by computer at Military Entrance Processing Stations (MEPS) or in written form at a satellite location called a Military Entrance Test site (MET). The subtests include General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Mathematics Knowledge, Electronics Information, Automotive and Shop Information, Mechanical Comprehension, Assembling Objects, and Verbal Expression. Navy applicants also complete a Coding Speed test. Over the years various measures of perceptual speed and accuracy and spatial ability have been utilized and discarded, but not lost. After considerable research on predicting job performance, measures such as Coding Speed and Assembling Objects have been found to be highly predictive of job performance (Held, Carretta, & Rumsey, 2014). For a comprehensive review of the research which has been conducted on performance measurement and classification efficacy in the military, the reader is referred to the article “Military Enlistment Selection and Classification: Moving Forward” (Rumsey & Arabian, 2014).

## Language acquisition

Another arena in which measures of aptitude have found a home is in the acquisition of second languages. There have existed for a number of years language aptitude tests that have been used to predict rate of acquisition of a second language at earlier stages in life (Carroll, 1981, 1985, 1993; Carroll & Sapon, 1959; Ehrman & Oxford, 1995; Hyltenstam & Abrahamsson, 2003). However, while there has been much less systematic investigation of adult learners who have achieved high-level proficiency in a second language, a few studies have suggested that highly successful learners possess a particular aptitude, or talent, for language learning (Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; Ioup, Boustagui, El Tigi, & Moselle, 1994). Furthermore, some researchers have argued that this high-level

aptitude is distinct from the more traditional conceptualizations of language aptitude (Linck et al., 2013; Schneiderman & Gesmarais, 1988). Linck et al. (2013) examined the ability of the High-Level Language Aptitude Battery to distinguish very successful language learners from other individuals. The battery consists of 11 cognitive and perceptual measures: Running Memory Span, Antisaccade Test, Stroop Test, Task Switching Numbers Test, Letter Span, Non-Word Span Test, Paired Associates, Available Long-Term Memory Synonym Test, Serial Reaction Time Test, and Phonemic Discrimination and Phonemic Categorization Tests. Results from the various analyses conducted resulted in a 70% classification accuracy rate (Linck et al., 2013). Measures of phonological short-term memory (Letter Span and Non-Word Span), implicit learning (Serial Reaction Time), and associative memory (Paired Associates) proved to be robust predictors of high attainment (Linck et al., 2013). On the other hand, executive function measures (Switch Cost component of Task Switching) did not show the predicted positive relationship with high-attainment outcomes with respect to listening analysis, suggesting that individuals with greater flexibility in mental shifting were less likely to be high attainers in listening versus reading. Based on these results, being capable of focusing one's attention squarely on the second language and preventing both controlled and uncontrolled attentional shifts to the first language appear important to developing high-level second language proficiency.

## Achievement testing

### *Achievement testing in the public schools*

For the past 15 years, since the enactment of the [No Child Left Behind Act of 2001](#) (NCLB), achievement testing in the nation's public schools has made major headlines in nearly every form of media because of the far-reaching consequences of its use. For example, a recent article in the Herald-Sun, a local paper in Durham, NC, discussed the state senate-proposed budget and its impact on teacher raises, which would reward teachers based on students' achievement testing scores. The journalist wrote that: "Senators also would spend more on pilot programs to reward teachers for superior performance, including \$10 million for third-grade reading teachers whose students make the most progress in standardized testing in each local district and statewide" ([Robertson, 2016](#)).

In the larger educational community, the conclusion that achievement testing results have now become synonymous not only with the measurement of student learning but also with the quality of teacher performance in the classroom (an addition to the requirement for teachers in core content areas to be "highly qualified in each subject taught") continues to be hotly debated issue in current educational and sociopolitical realms. Studies conducted in Florida and North Carolina suggested that some school administrators responded to test-driven accountability pressure by assigning their strongest teachers to the grades that were to be tested in order to boost their schools' performance ([Cohen-Vogel, 2011](#)). In addition, a nationwide

survey by the National Education Association found that teachers residing in states where standardized tests were required spent 29% of their work time on preparing students to take the tests, and very little time using test results to improve instruction ([National Education Association, 2013](#)).

While NCLB's test-driven accountability policies original intent was to lessen the gaps in academic achievement, particularly with respect to poor and minority students and their more advantaged peers in the areas of math and reading, the result was to narrow curricular offerings and become more focused on achievement tests. By 2010, however, 38% of schools were failing to make "adequate yearly progress (AYP)," up from 29% in 2006. As a consequence, the new federal administration offered a reprieve to many schools through a series of waivers, especially in the area of teacher evaluation—the biggest struggle for states. In exchange for the now unachieved aim of getting all students to proficiency, the states had to agree to set standards aimed at preparing students for higher education and the workforce ([www.edweek.org/ew/section/multimedia/no-child-left-behind-overview-definition-summary.html](http://www.edweek.org/ew/section/multimedia/no-child-left-behind-overview-definition-summary.html)). With the failure of NCLB to meet its targeted date of full implementation of its stated objectives by the 2013–14 school year, the congress enacted a new law, the [Every Student Succeeds Act \(ESSA\) in December 2015](#). In May 2016 regulations implementing this law were published in the Federal Register and remained open to comment until August 2016.

Assessing students with disabilities further complicated the matter when students' proficiency did not meet required levels at the specified time points (grades 3, 8, and at least once in high school). While the measurement of growth is applauded by nearly every organization that advocates for students with disabilities, these students on average were still performing well below their nondisabled peers ([Buzick & Laitusis, 2010](#)). And while local school districts were given the option of providing modified or alternative testing measures for these students, they were limited to counting up to 1% of the total student population as proficient for AYP based on scores obtained from alternate assessments. Only 2% of students taking modified assessments could be counted as proficient.

Tracking of these students from year to year becomes problematic as the scores from the various test forms may be impossible to combine in some longitudinal models. Further, some states exclude students taking alternate forms of the assessment tools as they set about measuring growth. Because students with disabilities are not a homogenous group, their numbers in specific classifications can be quite small, and they may move from one classification to another in the course of their education. While using measures of academic growth from annual state standards-based assessments is supported by the educational policy community and disability advocates, it is clear that an "Empirical understanding of growth measures for students with disabilities and the consequences of their use will contribute to more meaningful and appropriate decision making" ([Buzick & Laitusis, 2010](#), p. 543). In other words, the problems cited above will continue to need to be solved even with the most recent enactment of the Elementary and Secondary Education Act originally put into law in 1965.

## Achievement testing in clinical practice

Achievement tests can be classified as falling into three main categories, namely, group administered, individually administered, and modality-specific tests. Those generally under the domain of evaluators outside of both public and private educational institutions fall in the latter two categories, individually administered and/or modality-specific tests. For that reason, we will focus on those measures most commonly used in both psychological and neuropsychological assessments. In previous work in this area, Katz and Slomka (1990) provided a list of commonly used achievement measures in place at that time, which included the Woodcock–Johnson Psychoeducational Battery (WJ) (Woodcock, 1977), Wide Range Achievement Test-Revised (WRAT-R) (Jastak & Wilkinson, 1984), Peabody Individual Achievement Test-Revised (Markwardt, 1989), Kauffman Test of Educational Achievement (Kaufman & Kaufman, 2004), and Basic Achievement Skills Individual Screener (BASIS) (Psychological Corporation, 1983). Since that time, several of the tests have been updated and renormed. For example, the WJ has been revised several times and is now in the fourth edition (Woodcock, Shrank, McGrew, & Mather, 2005), and WRAT-R is also in the fourth edition. There has also been significant development and expansion of tests such as the Wechsler Individual Achievement Test (WIAT), which is now in the third edition (Pearson Inc., 2009). However, the scoring systems associated with the assessment of academic achievement remain essentially the same as those detailed by Katz and Slomka, and we refer the reader to their work for a more comprehensive discussion of scoring alternatives. An updated list of achievement tests is provided in Table 5.1.

While test developers continue to provide grade-equivalent scores, age-based scores, and percentile scores, standard scores are considered the more accurate and precise means of reporting test results (Katz & Slomka, 1990), which include  $z$  scores,  $t$  scores, and occasionally stanine scores. What remains equally critical is the norming group from which the standard scores are obtained, whether they be age or level of education based. This became a significant issue in a recent case where a client was seeking accommodations on step 1 of the USMLE conducted by the National Board of Medical Examiners (NBME). Upon review of the applicant's request for reasonable accommodations including additional time for testing, the NBME decided it was necessary to provide both the age-based and grade-based norms on the Woodcock–Johnson III NU Tests of Achievement to satisfy requirements under the Americans With Disabilities Act of 1990, ADA. The result was that the inclusion of both types of normative scores ensured the applicant received the appropriate accommodations, which had previously been questioned.

## Updated research on commonly used measures of achievement

In addition to the earlier work cited by Katz and Slomka (1990), Mather and Abu-Hamour (2013) provided an excellent comprehensive review regarding the

**Table 5.1** Commonly used achievement tests

<b>Group administered achievement tests</b>	
TerraNova 2/California Achievement Test—2nd ed. (CAT 6)	CTB/McGraw-Hill (2005). <i>TerraNova: second edition (CAT 6)</i> . Monterey, CA: Author.
Iowa Test of Basic Skills (ITBS)	Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2007). <i>Iowa test of basic skills (ITBS)</i> . Itasca, IL: Riverside.
Stanford Achievement Test—10th ed. (SAT-10)	Pearson (2003). <i>Stanford achievement test: 10th edition (SAT)</i> . Boston, MA: Author.
<b>Individually administered achievement tests</b>	
Basic Achievement Skills Inventory (BASI)	Bardos, A. (2004). <i>Manual for the BASI basic achievement skills inventory comprehensive test (BASI)</i> . San Antonio, TX: Pearson.
Kaufman Test of Educational Achievement—3rd ed. (KTEA-3)	Kaufman, A. S., & Kaufman, N. L. (2014). <i>Kaufman test of educational achievement: Third edition (KTEA-3)</i> . Bloomington, MN: Pearson.
Wide Range Achievement Test—5th ed. (WRAT-5)	Wilkinson, G. S., Robertson, G. J. (2017). <i>Wide range achievement test: Fifth edition (WRAT-5)</i> . Wilmington, DE: Jastak Associates.
Wechsler Individual Achievement Test—3rd ed. (WIAT-III)	Wechsler, D. (2009). <i>The Wechsler individual achievement test: Third edition (WIAT-III)</i> . San Antonio, TX: Pearson.
Woodcock—Johnson Tests of Achievement—4th ed. (WJ-IV-ACH)	Schrank, F. A., McGrew, K. S., Mather, N., & Woodcock, R. W. (2014). <i>Woodcock—Johnson tests of achievement: Fourth edition (WJ-IV-ACH)</i> . Rolling Meadows, IL: Riverside.
<b>Modality-specific achievement tests</b>	
Reading	
Classroom Reading Inventory	Weelock, W. H., Campbell, C. J., & Silvaroli, N. J. (2012). <i>Classroom reading inventory: One hundred twelfth edition</i> . New York, NY: McGraw-Hill.
Gates—MacGinitie Reading Tests (GMRT-4)	MacGinitie, W. H., MacGinitie, R. K., Maria, K. & Dreyer, L. G. (2000). <i>Gates—MacGinitie reading test: Fourth edition (GMRT-4)</i> . Itasca, IL: Riverside.
Gray Oral Reading Tests—5th ed. (GORT-5)	Wiederholt, J. L., Bryant, B. R. (2012). <i>Gray oral reading tests: Fifth edition (GORT-5)</i> . Austin, TX: PRO-ED.

(Continued)

**Table 5.1** (Continued)

Modality-specific achievement tests	
Nelson–Denny Reading Test	Brown, J. L., Fishco, V. V., & Hanna, G. (1993). <i>Nelson–Denny reading test</i> . Chicago: Riverside.
Woodcock Reading Mastery Tests—3rd ed. (WRMT-III)	Woodcock, R. W. (2011). <i>Woodcock reading mastery tests: Third edition (WRMT-III)</i> . Bloomington, MN: Pearson American Guidance Service.
Mathematics	
KeyMath-3 Diagnostic Assessment	Connolly, A. J. (2008). <i>KeyMath-3 diagnostic assessment. A diagnostic inventory of essential mathematics</i> . San Antonio, TX: Pearson.
Test of Mathematical Abilities—3rd ed. (TOMA-3)	Brown, V. L., Cronin, M. E., & Bryant, D. E. (2012). <i>Test of mathematical abilities: Third edition (TOMA-3)</i> . Austin, TX: PRO-ED.
Language	
Test of Written Language—4th ed. (TOWL-4)	Hammill, D. D., Larsen, S. C. (2009). <i>Test of written language: Fourth edition (TOWL-4)</i> . Austin, TX: PRO-ED.
Boston Diagnostic Aphasia Examination—3rd ed. (BDAE-3)	Goodglass, H., Kaplan, E., & Barresi, B. (2000). <i>The Boston diagnostic aphasia examination: Third edition (BDAE-3)</i> . Philadelphia: Lippincott Williams & Wilkins.
Multilingual Aphasia Exam (MAE)	Benton, A. L., Hamsher, K., De, S., Silvan, A. B. (1994). <i>Multilingual aphasia examination (MAE)</i> . Iowa City, IA: AJA Associates.
Neuropsychological Assessment—2nd ed. (NEPSY-II)	Korkman, M., Kirk, U., & Kemp, S. (2007). <i>Neuropsychological assessment: Second edition (NEPSY-II)</i> . San Antonio, TX: Harcourt Assessment.

individual assessment of academic achievement, which adds to the clinician's database to a significant degree. In their chapter, the authors provide a discussion of the various formal and informal measures of achievement. They briefly review studies that focused on achievement measures with a number of individuals with specific disabilities including dyslexia and attention deficit hyperactivity disorder (ADHD), issues with English as a second language users, and those with sensory impairments. They then detail the measurement of specific modalities such as reading (phonological awareness, basic reading skills, reading fluency and comprehension), written language expression, and mathematics. They conclude their work by summarizing the purposes of individualized academic achievement (Mather & Abu-Hamour, 2013). Those purposes are to (1) establish present levels of achievement, (2) determine what the student can and cannot do, (3) pinpoint patterns of strengths and weaknesses, (4) identify ways to measure and monitor future academic

progress, and (5) determine specific educational needs (Mather & Abu-Hamour, 2013, p. 122). Their work and those purposes remain salient. Therefore our focus will be on a number of test reviews and research studies on the most widely used achievement measures, including those available in the areas of reading and mathematics, which are basic skills in the educational and vocational worlds.

## **Wide Range Achievement Test—fourth edition**

The Wide Range Achievement Test—fourth edition (WRAT4) by Wilkinson and Robertson, published in 2006, is the most recent of the WRAT measures which date back to 1946 (Wilkinson & Robertson, 2006). The original Word Reading, Spelling, and Math Computation subtests have been expanded in the WRAT4 to include a Sentence Comprehension subtest and a reading composite score. The test with alternate forms is designed to be used with individuals between the ages of 5 and 94; and while brief in nature, it provides a foundation from which to further assess academic skill competencies if necessary. For the seasoned clinician, the Spelling subtest, for example, is a useful first step in determining if there may be a deficit in decoding on account of phonological awareness or processing issues because of the phonemic and nonphonemic nature of the stimulus words. The Word Reading subtest again provides data with respect to decoding and printed word recognition. The Math computation subtest gives the opportunity to identify errors of inattention to detail (misread operational signs) or the misalignment of numbers as examples, as well as initially identifying skills or lack thereof in basic math operations to the level of basic algebraic equations without the use of a calculator.

Finally, the Sentence Comprehension subtest is very similar to the Passage Comprehension subtest on the Woodcock–Johnson III NU (Woodcock et al., 2005) where the individual must supply a key word using a cloze procedure (Taylor, 1953), that is, responding to a deleted word or words from a portion of a text that implicitly requires the ability to understand context and vocabulary. Obviously, the more verbal cognitive skills to which the individual has access, the better he or she will perform on such a measure of reading comprehension when only one word is missing from the sentence, regardless of its length. This has been our experience and has led us to caution other evaluators when relying solely on a single-word cloze format for assessing reading comprehension, as it tends to mask problems with fluency and working memory functions that are not uncovered in the evaluation process unless lengthier, more textbook-like, multiparagraph reading passages are administered as well. This issue will be addressed in more detail when we look specifically at reading comprehension measures.

The test itself has been used in numerous research studies explicitly to obtain a baseline for academic achievement across of variety of subject cohorts, and while there have been criticisms in the past regarding the concurrent validity the WRAT3 with respect to reading comprehension (Marby, 1995), there is no real basis for the criticism in the opinion of these authors, as the measure never purports to be a

measure of comprehension per se, which is a complex process of which word recognition is only one component. In the words of the test authors: “The WRAT4 is intended for use by those professionals who require an assessment of important fundamental academic skills. Such measures are valuable in initial evaluation of individuals referred for learning, behavior, or vocational difficulties. The results of WRAT4 by themselves are not intended to provide formal identification of learning or cognitive disorders” (Wilkinson & Robertson, 2006, p. 3).

In contrast to Marby’s (1995) criticisms of the WRAT3, a recent study (Sayegh, Arentoft, Thaler, Dean, & Thames, 2014) examined the Reading’s subtest construct validity as an educational quality measure (QEd). They found that performance on the WRAT4 significantly predicted neuropsychological test performance versus quality of education ratings while controlling for race-ethnicity and socioeconomic status. Previous work had found reading achievement to correlate with QEd indicators in a meta-analysis study (Hedges, Laine, & Greenwald, 1994) and that reading tests were a better predictor of cognitive performance than standard years of education (Manly, Jacobs, Touradji, Small, & Stern, 2002). Sayegh and associates (2014) concluded that while clinicians work to gather as much information as possible when making informed decisions, “in many cases (and settings) gathering comprehensive information about QEd may not be feasible, and a relatively quick, practical measure such as the WRAT4 may be especially useful” (p. 735).

## Woodcock–Johnson IV Tests of Achievement

The Woodcock–Johnson IV Tests of Achievement, WJ-IV (Houghton Mifflin Harcourt, 2014), which replaced the third edition, is an individually administered measure for ages 2 through over 90 years of age. It contains a series of subtests in the areas of reading, mathematics, and written language. The standard battery consists of 11 tests. There are three alternate and parallel forms although six tests are considered the core set and are required for calculating intra-achievement variations (Houghton Mifflin Harcourt, 2014). There are Fluency measures in sentence reading, math facts, writing, and word reading. Reading Recall, Number Matrices, and Word Reading Fluency are all new additions to the test. There are also 22 Cluster Scores that can be calculated but they require administration of the Extended Battery which includes nine additional tests. It is important to note that all of the Cluster Scores, with the exception of Math Problem Solving, Reading Rate, and Reading Comprehension-Extended, were also available with the Woodcock–Johnson III Normative Update (NU) Tests of Achievement (Woodcock et al., 2005) and its computerized scoring system.

There are seven reading clusters including Reading, Reading Fluency, Reading Rate, Reading Comprehension, Reading Comprehension-Extended, Broad Reading, and Basic Reading Skills. The four math clusters cover Mathematics, Math Problem Solving, Broad Mathematics, and Math Calculation Skills. The four written language clusters include Written Language, Broad Written Language,

Basic Writing Skills, Written Expression. These clusters are essentially the same as the written language clusters found in the third edition. In addition, there is a Phoneme–Grapheme Knowledge Cluster that covers some more basic aspects of writing. There are also several broad academic clusters that consist of the following: Academic Skills, Academic Applications, Academic Fluency, Academic Knowledge, Brief Achievement, and Broad Achievement.

In a review of the WJ-IV, [Villarreal \(2015\)](#) suggests that positive aspects of the test include that it was developed with a large, nationally representative sample. It has acceptable reliability and validity. The test materials are well structured and administration procedures are user-friendly. The clusters are aligned with reading, writing, and math categories listed in specific definitions of learning disabilities ([Individuals With Disabilities Education Improvement Act, 2004](#)), which is particularly important within evaluations of school-aged children. Like any test, however, there were also some changes which some may view as limitation with WJ-IV. Specifically, the sample of skills was too limited for comprehensive instructional planning. Many of the tests have inadequate floors for very young children. There are high correlations between clusters in related areas, suggesting some redundancy between clusters. There is a complete shift to an online scoring and data management system which may cause concerns about privacy and confidentiality. [Villarreal \(2015\)](#) concludes that despite some of the minor limitations, the WJ-IV is a strong test which meets its stated purpose.

## Wechsler Individual Achievement Test-III

The other most commonly used, comprehensive, achievement battery by practitioners is the WIAT-III ([Pearson Inc., 2009](#)). The authors of the WIAT-III have worked to make the test more comprehensive, cover all the areas required by the Disabilities Education Improvement Act of 2004 ([IDEA, 2004](#)), has several additions which improve its utility over the WIAT-II ([Pearson Inc., 2009](#)). The structure of the WIAT-III now has seven composite scores, each made up of several subtests. Specifically, the Oral Language composite consists of the Listening comprehension and Oral Expression subtests. The Total Reading Composite consists of Word Reading, Pseudoword Decoding, and Reading Comprehension subtests. The Basic Reading Composite is based on the Word Reading and Pseudoword Decoding subtests. There is also a Reading Comprehension and Fluency Composite which is based on the Reading Comprehension subtest, timed measures included on the Word Reading and Pseudoword subtest, and the Oral Reading Fluency subtest. The addition of the reading fluency measures improves an area lacking in the WIAT-II. The Written Expression Composite is made from the Alphabet/Writing Fluency, Spelling, Sentence Composition, and Essay Composition subtests. The Mathematics Composite includes the Math Problem Solving and Numerical Operations subtests, similar to the WIAT-II, but has added a Math Fluency Composite that consists of Math Fluency Addition, Subtraction, and Multiplication subtests ([Pearson Inc., 2009](#)).

In his review of the WIAT-III, [Burns \(2010\)](#) asks “What is the ‘Gold Standard’ for measuring Academic Achievement?” He points out that this third version of the WIAT now provides the required academic areas specified by the Disabilities Education Improvement Act of 2004 ([IDEA, 2004](#)) for identifying a learning disability with the measurement of eight areas of achievement similar to those found on the WJ-IV. While the time required to assess younger children from kindergarten until the third grade ranges between 35 and 94 minutes, respectively, it takes more than 100 minutes to assess students between the 4th and 12th grades, which is a limitation of the test ([Burns, 2010](#)). In addition, the Essay Composition subtest and the Sentence Combining and Sentence Building subtests are somewhat time-consuming to score, and the Essay Composition subtest in particular may take practice in terms of scoring the resultant product. On the other hand, there is a linking sample that provides a comparison of 116 children on the WIAT-III and the WISC-IV as reported by Wechsler in 2003. Burns concludes while it is likely that the WIAT-III will be useful in a battery of neuropsychological tests, “the cost and time to administer may be the most important stumbling blocks” (p. 236). But, when the reason for referral is centered on an academic issue, Burns believes that the WIAT-III is in an excellent choice for assessing school-aged children. Regardless, the choice of test is often based on one’s preference. For example, the first author prefers to use the WJ-IV, and the second author of this chapter prefers the new version of the WIAT-III. However, the two tests are arguably more comparable now than with prior versions.

## **Individual achievement test-revised—normative update (PIAT-R/NU)**

The *PIAT-R/NU* was updated and published by Pearson Inc. in 1998 with new standardization data based on a national sampling of over 3400 school children and young adults (5–22 years of age). However, this was really simply an addition of new normative data as there were no changes made to the content of the test. While it is a less time-consuming measure than the WIAT-III (administration time is 60 minutes), which may be an asset, it is also its main drawback if the clinician is interested in a more comprehensive evaluation of academic achievement. The instrument offers six subtests that include General Information, Reading Recognition, Reading Comprehension (choosing one of four pictures that best illustrates a sentence), Written Expression (the requirement to write a story about a picture), Mathematics (multiple-choice items that test knowledge and application of math concepts and facts), and Spelling (multiple-choice items that measure recognition of correct word spelling). Most items can be responded to by pointing, which may be perceived to be a less threatening response format for some children.

An in-depth study ([Ott, 2011](#)) compared the reading subtest scores from the WIAT-III and the PIAT-R/NU. While both were significantly correlated, the WIAT-III subtests yielded significantly lower scores than did the similar reading

subtests on the PIAT-R/NU. It was also not clear that the two tests measure the same construct, perhaps because the WIAT-III contains additional tests that include Oral Reading Fluency, Reading Skills, and Pseudoword Decoding. In addition, the Reading Comprehension subtest on the PIAT-R/NU differs from the comparable WIAT-III test because it is presented in a pictorial multiple-choice format and the examinee is required to select the correct choice from memory of what was read. For the WIAT-III, the reading passage remains in front of examinees when they are asked to orally answer questions about the passage. Finally, as the instrument is not intended to yield diagnostic information, it can serve as an initial screening instrument with respect to academic achievement in general and is useful when assessing individuals with limited expressive abilities ([Markwardt, 1997](#)). However, it should not be used when testing is needed to document a learning disability.

## **Measurement of underlying cognitive processes involved in reading achievement**

The measurement of reading achievement is a complex undertaking because it includes a series of separate but integrated processes. Deficits in word recognition, reading rate, reading fluency, decoding skills, vocabulary, or weaknesses in listening and language comprehension can underlie an impairment in the development of reading comprehension as an academic skill set. To date, extensive research studies have documented that problems in several distinct areas can, and do, contribute to differences between good and poor readers ([Swanson & Hsieh, 2009](#); [Wagner & Torgesen, 1987](#); [Wolf & Bowers, 1999](#)). As a result, a number of instruments have been developed to further explicate specific underlying deficits if problems in reading comprehension are found during the measurement of achievement. There are at least eight commonly used measures of phonological awareness itself. A list of these can be found in the [Mather and Abu-Hamour \(2013\)](#). In addition, the various test publishing companies have several others on the market at this time.

The particular measure with which we are most familiar is the Comprehensive Test of Phonological Processing (CTOPP) ([Wagner, Torgesen, & Rashotte, 1999](#)). The CTOPP is a measure of phonological awareness, phonological memory, and rapid naming. Difficulty in any of these areas is considered a common cause of reading disability. The first version of the test was developed for children between the ages of 5 and 6, but we are most familiar with the second version, available for individuals between the ages of 7 and 24. The second version consists of seven core subtests and six supplemental subtests. In our experience, it is necessary to administer all 12 subtests with adolescents and young adults as there appears to be a ceiling effect when just the core subtests are given. There is now a CTOPP-2 that contains normative data collected in 2008 and 2009 which was published in 2013 ([Wagner, Torgesen, Rashatte, & Person, 2013](#)). However, the basic instrument remains the same and results in five composite scores: Phonological Awareness; Phonological Memory; Rapid Symbolic Naming; Rapid Non-Symbolic Naming; and an Alternate Phonological Awareness Composite Score.

The CTOPP does a remarkable job of sorting out phonological deficits from rapid naming deficits. Rapid naming deficit is a factor important for reading skills that was identified early in the work of Denckla and Rudel (1976). Rapid naming skills have more recently become the focus of research and intervention studies that are concerned with identifying and diagnosing an underlying developmentally based learning disorder involving reading fluency and its impact on comprehension (Wolf & Bowers, 1999; Wolf & Katzir-Cohen, 2001; Wolf, 1986, 2016; Wolf, Miller, & Donnelly, 2000). This research has found that the majority of children with developmental reading disabilities start with weaknesses in naming speed, and then go on to develop problems in reading fluency. Thus the ability of the CTOPP to differentiate these various skills is crucial to better understand an individual's reading disorder.

The role of fluency is also an important factor in reading disorders. As discussed earlier, reading fluency has been added to several test batteries (e.g., WIAT-III and WJ-IV). This is at least in part due to the work of Wolf. Indeed, in recent years, Wolf and Katzir-Cohen (2001) explicated a new conceptualization for fluency. They propose that reading fluency is the product of a combination of the initial development of accuracy and automaticity in core reading that include the sublexical and lexical processes. These processes are perceptual, phonological, orthographic, and syntactic in nature, and important for integrating single word and connected text. Once fully developed, fluency relies on a combination of accuracy and rate in which the actual decoding is relatively effortless, oral reading is smooth with correct prosody, and attention can be focused on comprehension rather than the basic mechanics of reading fluency. Based on these considerations, various strategies have been formulated to increase reading fluency. There are several excellent papers (Alber-Morgan, 2006; Mastropieri, Leinart, & Scruggs, 1999) detailing research validated strategies, which can be incorporated into achievement findings and recommendations.

## **Reading comprehension measures: research and critiques**

As discussed previously, the Reading Comprehension subtests of the WJ-IV and the WIAT-III use a cloze format to assess reading. Earlier work investigating reliability and validity of the cloze procedure is extensive, but one paper in particular (Cunningham & Tierney, 1979) raised the possibility that it may be easier to either predict or retrieve an author's choice of words for a narrative rather than an informational selection. Cloze format tests may tap into word decoding skills to a somewhat greater degree than question–answering tests probably because the individual can rely on the gist of a passage or previous background knowledge of the subject when answering a series of typical comprehension questions with one word deletions in the stimulus sentence, which then skews the resulting measurement score. We found this to be true when testing students of various ages.

With respect to testing format, it is interesting to note that there has been a shift from a multiple-choice format to an open-ended format for reading comprehension questions with the most recent edition of the Gray Oral Reading Tests—fifth edition

(GORT-5; Wiederholt & Bryant, 2012). The GORT-5 can be used to evaluate children and young adults between the ages of 6 and 23 years 11 months. It was designed to measure oral reading abilities including rate, accuracy, fluency, and comprehension. It has demonstrated evidence of strong psychometric properties (Hall & Tannebaum, 2013; Wiederholt & Bryant, 2012). Convergent-related validity evidence came from comparison to several other previously developed reading tests, including the Nelson–Denny Reading Test (NDRT; Brown, Fishco, & Hanna, 1993). It should be noted, however, that the GORT-5 uses a method of miscue analysis which is highly detailed and measures multiple skills. Therefore increased training time may be needed for anyone new to this instrument compared to other tests (Hall & Tannebaum, 2013).

The other option to open-ended questions or cloze format tests in the clinical setting is the measurement of comprehension via multiple-choice questions. Ozuru, Briner, Kurby, and McNamara (2013) reported an interesting study with college students examining performance on multiple-choice and open-ended questions and how the quality of self-explanations and level of prior knowledge contributes to those performances. Among their findings was that students' levels of topic-specific knowledge were more strongly correlated with performance on multiple-choice questions than with performance on open-ended questions, and that the amount of active processing as measured by the quality of self-explanation was more positively correlated with the open-ended versions of the same comprehension questions. Thus the results suggested that the same comprehension questions assessed different aspects of comprehension, automatic, passive or more controlled active information processing, depending on whether multiple-choice or open-ended questions were used (Ozuru et al., 2013).

Results from Ozuru et al. study (2013) may help to explain some of the criticisms in the literature with respect to reading comprehension measures such as the NDRT (Brown et al., 1993). In this test, the examinee reads a passage and is required to answer multiple-choice questions. This is an excellent measure if one wants to better understand difficulties that may be experienced when reading a text book or taking standardized tests. However, there have been studies which show that examinees have correctly answered the multiple-choice questions on the NDRT without reading the passages (Coleman, Lindstrom, Nelson, Lindstrom, & Gregg, 2010; Ready, Chaudhry, Schatz, & Strazzullo, 2013). Specifically, this test may have differential validity based on individual differences in vocabulary, general fund of knowledge, and broad reading skills (Coleman et al., 2010). For example, students who have difficulties with reading skills may guess the correct answer by relying on a general fund of knowledge. In addition, the earlier forms E and F of the NDRT were also suspect because they were time limited measures which may also contribute to guessing for those with limited reading fluency. Forms G and H of the NDRT now allow for extended time testing as a result. There is still room for improvement, however, as guessing correct answers on this test could further be reduced by a more balanced item composition (e.g., literal and interpretive items), and using passages which the reader would be unlikely to have prior knowledge of (Ready et al., 2013). These criticisms aside, when searching for a reading

comprehension measure that most closely resembles textbook reading samples in terms of both length and complexity of text content and a measure for reading rate, there is no other currently available measure similar to the NDRT for evaluating young adults with college degrees or beyond.

In this regard, positive correlations have been observed between scores on the NDRT, and measures of college students' study of expository prose as well as MCAT and NBME test scores (Feldt, 1988; Haught & Walls, 2002b; Jackson & Brooks, 1985; Jackson, Dawson-Saunders, & Jackson, 1984). Jackson and Brooks (1985) found the NDRT to be a better predictor of academic performance than the MCAT reading score, which may reflect the skills necessary to rapidly read and comprehend text in medical school. As a result of their work with the NDRT and medical school populations, there are new norms available for the measure when assessing healthcare professionals (Haught & Walls, 2002a), which we have found to be extremely useful when evaluating medical students for underlying learning disabilities and/or ADHD.

## Current measures of mathematical achievement

Among the measures of achievement in the area of mathematics, those found in the Woodcock–Johnson Achievement Battery (Houghton Mifflin Harcourt, 2014) and the WRAT4 (Wilkinson & Robertson, 2006) are excellent screening measures. Steiner and Ashcraft (2012) have developed three brief assessments of math achievement, two which are drawn from the WRAT and one composed of noncopyrighted items, for use with college students when lengthy testing is not feasible. Highly relevant for the clinician is their discussion of the changes between the WRAT3 and WRAT4 arithmetic problems. The authors point out that the newer test has removed four of the five most difficult items including the algebra problem with two unknowns, the factoring problem with exponents, and the function problem. Eight new items of lesser difficulty have also been added, making the WRAT4 somewhat less challenging and effectively decreasing the overall difficulty level of the test, according to the authors (Steiner & Ashcraft, 2012).

Earlier, the WRAT3 Arithmetic subtest was used in a study by Passolunghi, Marzocchi, and Fiorillo (2005) to identify the presence of an underlying arithmetic learning disorder (ALD) in a group of children with ADHD compared with a group of children with an ALD only and a control group achieving at normal levels. Children were included in the arithmetic learning disabilities group if their score was less than 29, which is two standard deviations below the mean according to Italian norms on the arithmetic subtest of the WRAT3. According to the authors of the study, children in the ADHD and the ALD group did not present any specific reading deficit using an Italian version of the reading subtest. As part of a comprehensive neuropsychological evaluation, the students were required to complete a set of eight arithmetic word problems that had been adapted from a standardized Italian test of arithmetic word problems in use for primary schools. Their results

showed that when working with arithmetic word problems, the children with ADHD recalled significantly more irrelevant literal information and significantly less relevant information than the other two groups. The ALD group was significantly more impaired in the solution of problems containing irrelevant numerical information. Both groups showed an impairment on working memory tasks (Passolunghi et al., 2005). They concluded that in all of the cognitive tasks involved in arithmetic word problems (comprehending the problem, constructing a representation of it, planning, and monitoring single subgoals), executive functions (particularly inhibition mechanisms) may explain an impairment in arithmetic word problem solving. Thus they provided evidence that problem solving difficulties of children with ADHD and ALD are related to the inability to reduce the memory accessibility of nontarget and irrelevant information.

Given the results of Passolunghi et al. (2005) and Steiner and Ashcraft (2012), the KeyMath-3 Diagnostic Assessment (Connolly, 2007) may be a useful instrument when a more comprehensive assessment of math skills is required. Available in two parallel forms for use with individuals between the ages of 4.5 and 21 years, the test items are grouped into 10 subtests that represent three general math content areas: basic concepts, operations, and applications. It covers the following item content: Numeration, Algebra, Geometry, Measurement, Data Analysis and Probability; Mental Computation and Estimation; Addition and Subtraction; Multiplication and Division; Foundations of Problem Solving; and Applied Problem Solving. The instrument reflects the content and process standards of the National Council of Teachers of Mathematics (2000). In addition, there is an extensive computer generated report that specifically details the individual's current skills in each area of measurement. Rosli (2011) provides a more in-depth review of the KeyMath Test for those interested in this test.

Finally, Brendefur and colleagues (2015) are in the process of developing the Primary Math Assessment (PMA) which is based on a multiple gating system and described as "A Comprehensive Mathematical Assessment Tool to Improve Mathematics Intervention for At-Risk Students." The instrument is designed to identify K-2 students at risk for poor math outcomes across six dimensions of math that are closely aligned with the common core math standards and include number sequencing, operations (number facts), contextual problems, relational thinking, measurement, and spatial reasoning (Brendefur et al., 2015). As early intervention is a key in the education of students at risk for failure, the instrument has the potential for being a significant addition to the field of assessment with younger children in the area of mathematics, once it is more fully validated.

## **Aptitude and achievement testing in the 21st century: comments and conclusions**

In this chapter we have focused on salient aspects of both aptitude and achievement testing and their implications for the clinician in current practice. We have offered

insights from our clinical experiences with the various measures as we deemed appropriate and ask that the readers consider their relevance to their own practices should they be helpful. What is clear to us from our review of the most common testing materials is that the normative samples have been updated with respect to the most recent population data since the publication of the third edition of this Handbook (Goldstein & Hersen, 1990).

It is interesting to note that in the area of reading in particular, the instruments now available for assessing the underlying components of reading comprehension are based on solid empirical evidence derived from studies conducted in the late 1990s and early 2000s, some of which were spurred on by political and social pressures and the resultant federal legislation at the time. In some regards this is similar to the development of aptitude measures, which came out of the Second World War and the need for skilled technicians in the war effort, to a large degree. Likewise, it seems highly probable that work in the area of measurement of mathematical skills and the learning process based on a knowledge of those skills or lack thereof will continue as the pressures for science, technology, and mathematical competence continue to be in the forefront of educational policies and socioeconomic realities.

It is our view that standardized measurements of achievement and validated measures of aptitude will continue to fill vital and functional roles as we attempt to promote high levels of learning for all students in the culturally diverse and technology enriched environment which currently exists. The assessment of both achievement and specific aptitudes may well help to steer individuals into successful careers in this rapidly developing world of technological and informational systems that will drive the global economy. When assessments are conducted to help guide improvement in instruction and in learning, and attention is paid to targeting an expanded set of competencies for which students are prepared, then we can say that they retain their value and merit continued study and refinement.

## References

- Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, 30, 481–509.
- ACT. (2007). *The ACT technical manual*. Iowa City, IA: Iowa University Press.
- Alber-Morgan, S. R. (2006). Ten ways to enhance the effectiveness of repeated readings. *Journal of Early and Intensive Behavior Intervention*, 3(3), 273–279.
- Americans With Disabilities Act. (1990). Retrieved from <<https://www.eeoc.gov/eeoc/history/35th/thelaw/ada.html>>.
- Anastasi, A. (1984). Aptitude and achievement tests: The curious case of the indestructible strawberries. In B. S. Plake (Ed.), *Social and technical issues in testing: implications for test construction and usage. Buros—Nebraska symposium on measurement and testing* (pp. 129–140).
- Bennett, G. K. (1940). *Manual of the test of mechanical comprehension, form AA*. San Antonio, TX: Psychological Corporation.

- Bennett, G. K. (1994). *Manual for the Bennett mechanical comprehension test, form S and T* (2nd ed). San Antonio, TX: Psychological Corporation.
- Bennett, G. K. (2008). *Bennett mechanical comprehension test: Administration manual*. San Antonio, TX: NCS Pearson, Inc.
- Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1947). *A manual for the differential aptitude Tests*. New York: Psychological Corporation.
- Brendefur, J. L., Johnson, E. S., Thiede, K. W., Smith, E. V., Strother, S., Severson, H. H., ... Beaulieu, J. (2015). Developing a comprehensive mathematical assessment tool to improve mathematics intervention for at-risk students. *Journal for Research in Learning Disabilities*, 2(2), 65–90.
- Bridgman, B., Pollack, J., & Burton, N. (2004). Understanding what SAT reasoning test scores add to high school grades: A straightforward approach. In T. C. Board (Ed.), *ETS Research Report Series* (Vol. 2004, pp. i–20). New York, NY: The College Board.
- Brown, J. I., Fishco, V. V., & Hanna, G. S. (1993). *Nelson–Denny reading test, forms G and H*. Austin, TX: Pro-Ed.
- Burns, T. G. (2010). Wechsler individual achievement test-III: What is the ‘Gold Standard’ for measuring academic achievement? *Applied Neuropsychology*, 17, 234–236.
- Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher*, 39(7), 537–544.
- Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. G. Dilla (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–118). Rowley, MA: Newbury House.
- Carroll, J. B. (1985). Second-language abilities. In R. J. Sternberg (Ed.), *Human abilities: An information-processing approach* (pp. 83–103). New York: W.H. Freeman.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test*. New York: Psychological Corporation.
- Cohen-Vogel, L. (2011). “Staffing to the test”: Are today’s school personnel practices evidence based? *Educational Evaluation and Policy Analysis*, 33(4), 483–505.
- Coleman, C., Lindstrom, J., Nelson, J., Lindstrom, W., & Gregg, K. N. (2010). Passageless comprehension on the Nelson–Denny reading test: Well above chance for university students. *Journal of Learning Disabilities*, 43(3), 244–249.
- Collin, V. T., Violato, C., & Hecker, K. (2009). Aptitude, achievement and competence in medicine: A latent variable path model. *Advances in Health Sciences Education*, 14(3), 335–366.
- Connolly, A. S. (2007). *KeyMath-3 diagnostic assessment: Manual forms A and B*. Minneapolis, MN: Pearson.
- Cunningham, J. W., & Tierney, R. J. (1979). Evaluating cloze as a measure of learning from reading. *Journal of Reading Behavior*, 11(3), 287–292.
- DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22(4), 499–533.
- Denckla, M. B., & Rudel, R. G. (1976). Rapid “automatized” naming (R.A.N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia*, 14(4), 471–479.
- Ehrman, M. E., & Oxford, R. L. (1995). Cognition plus: Correlates of language learning success. *Modern Language Journal*, 79(1), 67–89.
- Every Student Succeeds Act, Pub. L. No. 114-95, 20USC6301Stat. (2015). Retrieved from <<https://all4ed.org/essa/>>.

- Feldt, R. C. (1988). Predicting academic performance: Nelson—Denny reading test and measures of college students' study of expository prose. *Psychological Reports*, 63(2), 579–582.
- Goldstein, G., & Hersen, M. (1990). *Handbook of psychological assessment* (2nd ed.). Elmsford, NY: Pergamon Press.
- Hall, A. H., & Tannebaum, R. P. (2013). Test review: J. L. Wiederholt & B. R. Bryant (2012). Gray oral reading tests—fifth edition (GORT-5). Austin, TX: Pro-Ed. *Journal of Psychoeducational Assessment*, 31, 516–520.
- Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, 8(1), 1–51.
- Halpern, D. F., & Butler, H. A. (2013). Assessment in higher education: Admissions and outcomes. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 319–336). Washington, DC: American Psychological Association.
- Haught, P. A., & Walls, R. T. (2002a). Adult learners: New norms on the Nelson—Denny reading test for healthcare professionals. *Reading Psychology*, 23(3), 217–238.
- Haught, P. A., & Walls, R. T. (2002b). *Adult trainers: Relationship of reading, MCAT, USMLE step I test results for medical students*. Paper presented at the Annual meeting of the American educational research association, New Orleans, LA.
- Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). An exchange: Part I. Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, 23(3), 5–14.
- Held, J. D., Carretta, T. R., & Rumsey, M. G. (2014). Evaluation of tests of perceptual speed/accuracy and spatial ability for use in military occupational classification. *Military Psychology*, 26(3), 199–220.
- Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist*, 61, 845–859.
- Houghton Mifflin Harcourt. (2014). *Woodcock—Johnson IV tests of achievement*. Rolling Meadows, IL: Houghton Mifflin Harcourt.
- Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in second language acquisition. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 539–588).
- IDEA. (2004). Individuals with Disabilities Education Improvement Act.
- Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, 16(1), 73–98.
- Jackson, E. W., Dawson-Saunders, R., & Jackson, J. E. (1984). The predictive validity of the Nelson—Denny reading test for scores on the reading subtest of the MCAT. *The Advisor*, 5, 7–11.
- Jackson, J. R., & Brooks, C. M. (1985). Relationships among the MCAT reading subtest, Nelson—Denny reading test, and medical school achievement. *Journal of Medical Education*, 60(6), 478–480.
- Jastak, S., & Wilkinson, G. S. (1984). *Wide range achievement test-revised*. Wilmington, DE: Jastak Associates.
- Katz, L. J., Beers, S. R., Geckle, M., & Goldstein, G. (1989). Clinical use of the career ability placement survey vs. the GATB with persons having psychiatric disabilities. *Journal of Applied Rehabilitation Counseling*, 20(1), 13–19.

- Katz, L. J., & Slomka, G. T. (1990). Achievement testing. In G. Goldstein, & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd ed, pp. 123–147). Elmsford, NY: Pergamon Press.
- Kaufman, A. S., & Kaufman, N. L. (2004). Kaufman Test of Educational Achievement—Second Edition (KTEA-II). Circle Pines, MN: American Guidance Service.
- Knapp, L., Knapp, R. R., & Knapp-Lee, L. (1977). *CAPS: Career ability placement survey—examiner's manual*. San Diego, CA: EDITS.
- Knapp, L., Knapp, R. R., Strnad, L. S., & Michael, W. B. (1978). Comparative validity of the Career Ability Placement Survey (CAPS) and the General Aptitude Test Battery (GATB) for predicting high school course marks. *Educational and Psychological Measurement*, 38(4), 1053–1056.
- Kobrin, J. L., Carmase, W. J., & Milowski, G. B. (2002). The utility of the SAT I and SAT II for admission decisions in California and the nation. In T. C. Board (Ed.), *ETS Research Report Series* (pp. 1–28). New York, NY: The College Board.
- Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., ... Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, 63(3), 530–566.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437–448.
- Manly, J. J., Jacobs, D. M., Touradji, P., Small, S. A., & Stern, Y. (2002). Reading level attenuates differences in neuropsychological test performance between African American and White elders. *Journal of the International Neuropsychological Society*, 8 (3), 341–348.
- Marby, L. (1995). Review of the wide range achievement test—third edition. In J. C. Conoley, & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 1108–1110). Lincoln, NE: Buros Institute.
- Markwardt, F. C. (1989). *Peabody individual achievement test-revised*. Circle Pines, MN: American Guidance Service.
- Markwardt, F. C., Jr. (1997). *Peabody individual achievement test-revised—normative update*. Circle Pines, MN: American Guidance Service.
- Mastropieri, M. A., Leinart, A., & Scruggs, T. E. (1999). Strategies to increase reading fluency. *Intervention in School and Clinic*, 34, 278–283.
- Mather, N., & Abu-Hamour, B. (2013). Individual assessment of academic achievement. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education* (Vol. 3, pp. 101–128). Washington, DC: American Psychological Association.
- National Council of Teachers of Mathematics. (2000). *Principles and standards of school mathematics*. National Council of Teachers of Mathematics.
- National Education Association. (2013). *Member survey on standardized testing*. Washington, DC: Author.
- No Child Left Behind Act. 107–110 C.F.R. (2001). Retrieved from <<https://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf>>.
- Ott, L. M. (2011). Comparison of the reading subtests of the Wechsler Individual Achievement Test—third edition and the Peabody Individual Achievement Test-Revised/NormativeUpdate. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 71(8-A), 2847.

- Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(3), 215–227.
- Passolunghi, M. C., Marzocchi, G. M., & Fiorillo, F. (2005). Selective effect of inhibition of literal or numerical irrelevant information in children with attention deficit hyperactivity disorder (ADHD) or arithmetic learning disorder (ALD). *Developmental Neuropsychology*, 28(3), 731–753.
- Patterson, B. F., Mattern, K. D., & Kobrin, J. L. (2009). *Validity of the SAT for predicting FYGPA: 2007 SAT validity sample*. ETS Research Report Series. New York: The College Board.
- Pearson Inc. (2009). *Wechsler individual achievement test—third edition*. San Antonio, TX: Pearson Inc.
- Peterson, J. J. (1983). *The Iowa testing progress: The first fifty years*. Iowa City: University of Iowa Press.
- Psychological Corporation. (1983). *Basic achievement skills individual screener*. San Antonio, TX: Author.
- Ramsay Corporation. (2004). *Mechanical aptitude test—form MAT-4 (Online)*. Pittsburgh, PA: Ramsay Corporation.
- Ready, R. E., Chaudhry, M. F., Schatz, K. C., & Strazzullo, S. (2013). Passageless administration of the Nelson–Denny reading comprehension test: Associations with IQ and reading skills. *Journal of Learning Disabilities*, 46(4), 377–384.
- Robertson, O. D. (2016). Senate budget heavier on teacher raises, policy change. *The Herald Sun*.
- Rosli, R. (2011). Test review: A. S. Connolly KeyMath-3 diagnostic assessment: Manual forms A and B. Minneapolis, MN: Pearson; 2007. *Journal of Psychoeducational Assessment* (29, pp. 94–97).
- Rumsey, M. G., & Arabian, J. M. (2014). Military enlistment selection and classification: Moving forward. *Military Psychology*, 26(3), 221–251.
- Sayegh, P., Arentoft, A., Thaler, N. S., Dean, A. C., & Thames, A. D. (2014). Quality of education predicts performance on the Wide Range Achievement Test—4th edition. Word Reading Subtest. *Archives of Clinical Neuropsychology*, 29(8), 731–736.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262–274.
- Schneiderman, E. L., & Gesmarais, C. (1988). The talented language learner: Some preliminary findings. *Second Language Research*, 4(2), 91–109.
- Silver, C. H., Blackburn, L. B., Arffa, S., Barth, J. T., Bush, S. S., Koffler, S. P., & Elliott, R. W. (2006). The importance of neuropsychological assessment for the evaluation of childhood learning disorders: NAN Policy and Planning Committee. *Archives of Clinical Neuropsychology*, 21(7), 741–744.
- Silzer, R., & Church, A. H. (2009). The pearls and perils of identifying potential. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2(4), 377–412.
- Steiner, E. T., & Ashcraft, M. H. (2012). Three brief assessments of math achievement. *Behavior Research Methods*, 44(4), 1101–1107.
- Stemler, S. E., & Sternberg, R. J. (2013). The assessment of aptitude. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 3*.

- Testing and assessment in school psychology and education* (Vol. 3, pp. 281–296). Washington, DC: American Psychological Association.
- Swanson, H. L., & Hsieh, C.-J. (2009). Reading disabilities in adults: A selective meta-analysis of the literature. *Review of Educational Research*, 79(4), 1362–1390.
- Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415–433.
- US Department of Defense. (1984). *Test manual for the armed services vocational aptitude battery*. North Chicago, IL: US Military Entrance Processing Command.
- Villarreal, V. (2015). Review of Woodcock–Johnson IV tests of achievement. *Journal of Psychoeducational Assessment*, 33(4), 391–398.
- Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101(2), 192–212.
- Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Person, N. S. (2013). *Comprehensive test of phonological processing—second edition (CTOPP-2)*. North Tonawanda, NY: MHS Assessments.
- Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive test of phonological processing examiner manual*. Austin, TX: Pro-Ed.
- Webb, R. M., Lubinski, D., & Benbow, C. P. (2002). Mathematically facile adolescents with math–science aspirations: New perspectives on their educational and vocational development. *Journal of Educational Psychology*, 94(4), 785–794.
- Wiederholt, J. L., & Bryant, B. R. (2012). *Gray oral reading tests—fifth edition (GORT-5)*. Austin, TX: Pro-Ed.
- Wiesen, J. P. (1977). *Test of mechanical aptitude*. Scarsdale, NY: APR Testing Services.
- Wilkinson, G. S., & Robertson, G. J. (2006). WRAT Wide range achievement test: Professional manual. Lutz, FL: Pearson.
- Wolf, M. (1986). Rapid alternating stimulus naming in the developmental dyslexias. *Brain and Language*, 27, 360–379.
- Wolf, M. (2016). New research on an old problem: A brief history of fluency. Retrieved from <<http://www.scholastic.com/teachers/article/new-research-old-problem-brief-history-fluency>>.
- Wolf, M., & Bowers, P. (1999). The “Double-Deficit Hypothesis” for the developmental dyslexias. *Journal of Educational Psychology*, 91(3), 1–24.
- Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, 5, 211–238.
- Wolf, M., Miller, L., & Donnelly, K. (2000). The Retrieval, Automaticity, Vocabulary Elaboration, Orthography (RAVE-O): A comprehensive fluency-based reading intervention program. *Journal of Learning Disabilities*, 33(4), 375–386.
- Woodcock, R. W. (1977). *Woodcock–Johnson psychoeducational battery: Technical report*. Allen, TX: DLM Teaching Resources.
- Woodcock, R. W., Shrank, F. A., McGrew, K. S., & Mather, N. (2005). *Woodcock-Johnson III normative update*. Retrieved from Rolling Meadows, IL.

# Interest inventories

6

Jo-Ida C. Hansen

Center for Interest Measurement Research, Department of Psychology, University of Minnesota, Minneapolis, MN, United States

## Introduction

Interests, particularly vocational interests but also leisure interests, have been studied in vocational psychology for more than 100 years. During World War I, psychologists were recruited by the US government to develop techniques for assessing intelligence and personality for military personnel selection. Many of these same psychologists pursued the assessment of interests after the war as they found their way to employment in academic research institutions.

The importance of an individual's interests in job selection was first recognized by educators in the 1900s and shortly thereafter by industry. Early theorists in the field, such as Parsons (1909), hypothesized that occupational adjustment was enhanced if an individual's characteristics and interests matched the requirements of the occupation. As Strong (1943) noted in "Vocational Interests of Men and Women," interests provide additional information, not available from analyses of abilities or aptitudes, for making career decisions. Consideration of interests, along with abilities, values, and personality characteristics, provides a thorough evaluation of an individual that is superior to consideration of any trait in isolation.

Beyond the importance of interests for job selection, interests appear to have adaptive and motivational functions such that people who associate with others who have similar interests show higher levels of satisfaction and well-being than do those in less congruent interest work environments (Dik & Hansen, 2008). Similar to the positive relation between vocational interests and work satisfaction, research on leisure interests has shown the importance of congruence for satisfaction. For example, Ton and Hansen (2001) found that congruence of leisure interests and values for married couples was predictive of marital satisfaction, and interests and values together accounted for 28% of the variance in marital satisfaction.

The earliest method for assessing interests was *estimation*, accomplished by asking individuals to indicate how they felt about various activities. To improve the accuracy of their estimation, people were encouraged to *try-out* activities before making their estimates. However, try-out techniques for evaluating interests were time-consuming and costly; the search for a more economical assessment method

led to the development of interest *checklists* and *rating scales* (Kitson, 1925; Miner, 1922) and eventually to *interest inventories* that used statistical procedures to summarize an individual's responses to a series of items representing various activities and occupations.

## The earliest item pool

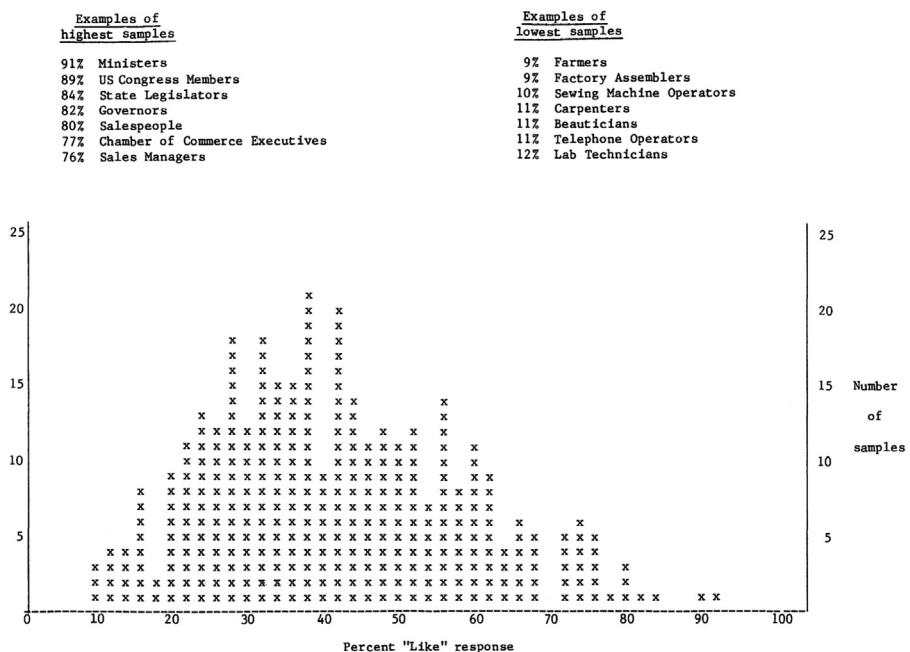
The first item pool of interest activities was accumulated by a seminar taught by Clarence S. Yoakum at Carnegie Institute of Technology in 1919. The 1000-item pool was developed using a *rational sampling approach* designed to represent the entire domain of interests. Over the years, statistical analyses were performed to determine the worth of each item, and numerous test authors used that original item pool as the foundation for development of their inventories [e.g., Occupational Interest Inventory (Freyd, 1922–1923); Interest Report Blank (Cowdery, 1926); General Interest Survey (Kornhauser, 1927); Vocational Interest Blank (Strong, 1927); Purdue Interest Report (Remmers, 1929); Interest Analysis Blank (Hubbard, 1930); and Minnesota Interest Inventory (Paterson, Elliott, Anderson, Toops, & Heidbreder, 1930)]. Among these inventories, the Vocational Interest Blank, now the Strong Interest Inventory (SII; Consulting Psychologists Press, 2004, 2012) is the only one to have remained in continuous use up to the present time. David P. Campbell and Charles H. Johansson both worked with the SII and used their knowledge about SII items and scale construction to develop the Campbell Interest and Skill Survey (CISS; Campbell, Hyne, & Nilsen, 1992) and the Career Assessment Inventory (CAI; Johansson, 1975).

## Characteristics of good items

Interest inventory items should be evaluated periodically because societal changes can make items obsolete as well as create the need for new items. Several qualities that contribute to the excellence of items, and ultimately to the excellence of an interest inventory, can be used to assess the value of each item.

First, items should differentiate among groups because the purpose of interest inventories is to distinguish people with similar interests from those with dissimilar interests. The item in Fig. 6.1, for example, has the power to spread 350 occupations over a wide range of response percentages. The lowest “Like” response rate for this item, *Making a Speech*, is 9% (meaning that few people in the sample answered “Like” to the item), and response rates range up to a high of 91% (meaning that the majority of the sample responded “Like”).

Samples of groups with similar interests should have similar item response rates, and clusters of groups with high or low response rates should make sense. In Fig. 6.1, for example, the samples composed of ministers, members of the US Congress, state legislators, governors, salespeople, and Chamber of Commerce



**Figure 6.1** Percent “Like” response to the item *Making a Speech* for the 350 Occupational Scales.

executives had high “Like” response rates of 77%–91%. Farmers, factory assemblers, sewing machine operators, carpenters, beauticians, and laboratory technicians, however, had low “Like” response rates to the same item. Those clusters of high response rate and low response rate samples are intuitively satisfying and illustrate the item’s content validity; one expects ministers and politicians, for example, to enjoy making a speech.

Items also should be sex-fair; no item should suggest that any occupation or activity is more appropriate for one sex than the other. For example, an item such as *Stewardess* that appeared on early forms of the Strong Vocational Interest Blank, was changed to *Flight Attendant* on the SII. In addition to sex-fair items, all interpretive and instructional materials for interest inventories should be sex-fair.

To facilitate the adaptation of inventories for use with ethnic minorities or for international use, interest items should be unambiguous and culture-fair. Straightforward items also are more likely to have the same meaning for everyone taking the inventory regardless of cultural or occupational orientation, and they will be easier to translate into other languages.

All items should be revised periodically to ensure that they are current and familiar to the respondents. The face validity, as well as content validity, of an interest inventory is affected if the item pool contains obsolete items that are unfamiliar to the general population. On the other hand, as new occupations emerge and new technologies develop, new items should be generated to ensure that the entire domain of

interests is represented in the item pool. For example, recent additions to the SII item pool include items related to computers/technology: *computer technician, software developer, information technology, constructing an Internet website, using computer-aided design software, providing technical support, networking computer systems, installing computer software, doing data entry, managing a computer database, conducting a search on the Internet, and diagnosing computer problems.*

Finally, items should be easy to read. All materials that accompany interest inventories (e.g., instructions, profile, interpretive information) and the item pool itself should be easy to read to make the inventory useful for a wide educational and age range in the population.

## **Influence of vocational interest theories on inventory construction**

The earliest interest inventories were developed using the atheoretical, empirical method of contrast groups that is based on an assumption that people with similar interests can be clustered together and, at the same time, be differentiated from groups with dissimilar interests. Inventories that still incorporate this method of scale construction are the SII ([Consulting Psychologists Press, 2004, 2012](#)), the CAI ([Johansson, 1975, 1986](#)) and the CISS ([Campbell et al., 1992](#)).

Results from the early empirical investigations of interests later were used to develop hypotheses and theories about the structure of interests. [Roe \(1956\)](#) and [Holland \(1959\)](#), for example, used the factor analyses of Guilford and his colleagues ([Guilford, Christensen, Bond, & Sutton, 1954](#)), who found seven interest factors: (1) mechanical, (2) scientific, (3) social welfare, (4) aesthetic expression, (5) clerical, (6) business, and (7) outdoor work, to guide the development of their theories about interests. Holland's theory ([Holland, 1997](#)), in particular, has been very influential in the development of interest inventories. He proposed six vocational types—Realistic, Investigative, Artistic, Social, Enterprising, and Conventional—that capture higher order interests. Holland's six types have been incorporated into the profiles of all of the most popular interest inventories.

### ***Construction of interest inventory scales***

Construction of interest inventory scales is based on several assumptions:

1. First, a person can give informed responses of degree of interest (e.g., like, indifferent, dislike) to familiar activities and occupations.
2. Then, unfamiliar activities have the same factor structure as do familiar activities.
3. Therefore familiar activities and occupations can be used as items in interest inventories to identify unfamiliar occupational interests.

Early interest inventories typically featured either *homogeneous* or *heterogeneous* scales. Generally, heterogeneous scales are more valid for predictive uses of

interest inventories (e.g., predicting future job entry or college major), but homogeneous scales are more useful for providing parsimonious descriptions of the structure of a sample's interests (Hansen, 2016).

## **Homogeneous scale development**

One method of scale construction involves clustering together items based on internal consistency or homogeneous scaling. Items chosen in this manner have high intercorrelations. Empirical methods such as cluster or factor analyses can be used to identify the related items. The scales of the Vocational Interest Inventory (VII; Lunneborg, 1976), for example, were constructed using factor analysis. Scales for the Leisure Interest Questionnaire (LIQ; Hansen, 1998; Hansen & Scullard, 2002) were constructed using a series of cluster analyses to reduce the pool of items and then, factor analyses of the final item pool.

The development of assessment scales also may be based on rational selection of items; this method uses a theory to determine items appropriate for measuring the construct represented by each scale. For example, the General Occupational Themes (GOT) of the SII were rationally constructed using Holland's theoretical definition of the six vocational types to guide item selection (Campbell & Holland, 1972; Hansen & Johansson, 1972). Holland (1971) used his own theory of vocational types to select items for the Self-Directed Search (SDS).

## **Heterogeneous scale development**

The Occupational Scales of the CISS, the CAI, and the SII are composed of items with low intercorrelations and, therefore, are called heterogeneous scales. Heterogeneous scales often are atheoretical; in other words, the choice of items is based on empirical results rather than an underlying theory. The CISS, CAI, and the SII use the empirical method of contrast groups to select items for the Occupational Scales; this technique compares the item response rates of occupational criterion groups and contrast groups, representing the interests of people-in-general, to identify items that significantly differentiate the two samples.

### ***Current interest inventories***

One of the most recently developed commercial interest inventories is the CISS (Campbell, 1995). Other widely used inventories include the SDS (Holland, 1971, 1979, 1985a, 1994; Holland & Messer, 2013b); the SII (Consulting Psychologists Press, 2004, 2012; Hansen & Campbell, 1985), and the CAI (Johansson, 1975, 1986). In addition to these commercially available inventories, the Interest Item Pool (IIP) is an open access item pool, similar to the International Personality Item Pool (IPIP). The IIP website offers public domain vocational interest scales that

assess Holland's six types as well as 31 basic interests. The O\*NET also presents online materials that facilitate interest exploration. One leisure interest inventory has been developed but is available only from the author (LIQ; [Hansen, 1998](#)).

## Campbell Interest and Skill Survey

David Campbell, author of the CISS, describes the instrument as a product of 90 years of psychometric evolution influenced to a large extent by his own work with the SII in the 1960s, 1970s, and 1980s ([Campbell, 1995](#)). The CISS is unique among interest inventories in that the instrument is designed to assess not only an individual's interest in academic and occupational topics but also an individual's estimation of her or his skill in a wide range of occupational activities. The profile includes 98 scales on which two scores are provided—an interest score and a skill score. The CISS is available in three formats—Web-based administration, scoring and reporting; scoring and reports using desktop software; and a mail-in-scoring service that involves sending the completed answer sheet to the publisher, who does the processing and returns a profile by regular mail.

*Item pool and profile.* The item pool for the CISS includes 200 interest items and 120 items designed to assess self-reported skills. The response format for the interest items is a six point scale ranging from Strongly Like to Strongly Dislike. The skill items also have a six point response scale that includes self-evaluations of Expert, Good, Slightly above Average, Slightly Below Average, Poor, and None (have no skills in this area).

*Scales.* The CISS profile includes three types of scales: seven Orientation Scales, 29 Basic Scales, and 60 Occupational Scales. The Orientation Scales capture the major interest factors that have been identified through various statistical clustering procedures and include Influencing (business and politics), Organizing (managing and attention to detail), Helping (service and teaching), Creating (the arts and design), Analyzing (science and math), Producing (hands-on and mechanical), and Adventuring (physical activities and competition). The Orientation Scales are used as the organizational frame of reference for the CISS profile. Six of the seven Orientation Scales capture Holland's six types—Influencing on the CISS measures Holland's Enterprising type, Organizing is similar to Holland's Conventional, Helping reflects Holland's Social, Creating is Holland's Artistic, and Producing is Holland's Realistic.

The 29 Basic Scales were developed by clustering together homogeneous items in content areas such as Sales, Supervision, Adult Development, International Activities, Science, Woodworking, and Military/Law Enforcement. The Basic Scales are grouped on the profile under the Orientation with which they correlate most highly. For example, Mathematics and Science Basic Scales have their highest correlation with the Analyzing Orientation Scale and thus appear together on the profile. Likewise, Supervision, Financial Services, and Office Practices Basic Scales appear with the Organizing Orientation.

The CISS Occupational Scales were constructed using the empirical method of contrast groups originally refined for interest measurement by E.K. Strong, Jr. Successful, satisfied workers in each of 60 occupations were surveyed. Their responses to each of the CISS items were compared to the item responses of a general reference sample composed of employed workers from a variety of occupations. Items that substantially differentiated the occupational criterion sample from the general reference sample were selected for the occupation's scale.

The first step to determine the location of the Occupational Scales on the profile was to compute the mean score for each occupational criterion sample on the Orientation Scales. The occupation's highest Orientation score then was used to locate the Occupational Scale on the profile. For example, the Test Pilot and Ski Instructor criterion samples scored highest on the Adventuring Orientation and therefore the Occupational scales representing their interests are clustered with the Adventuring Orientation on the profile.

Two additional scales on the CISS profile are Academic Focus and Extraversion. The Academic Focus scale measures interest and confidence in academic pursuits especially science and the arts. The Extraversion scale measures interest and confidence in activities that require high levels of personal interaction.

*Norming and score report.* All of the scales on the CISS are normed on a general reference sample of women and men. The scales also are standardized with the result that the mean score for the General Reference Sample is about 50 and the standard deviation about 10. The sample used to norm the scales included 1790 female and 3435 male respondents from 65 occupational samples. The raw score means for the two samples were averaged to give the sexes equal weighting in the raw-score-to-standard-score conversion.

The CISS Report is a lengthy document that includes pages that report the Orientation and Basic Scale Interest and Skill scores. Additional pages summarize the scores for all of the Occupational Scales, as well as a repeat of the Basic Interest and Skill Scales, related to each of the seven Orientations. Additional pages report the Special Scales (Academic Focus and Extraversion) and procedural checks, followed by a two page summary of all the scales and scores.

In addition to presenting an Interest and a Skill score for each scale, the profile also includes a graph that plots the Interest and Skill scores to provide interpretive comments ranging from very low to very high. An interpretive bar, representing the middle 50% of scores for each criterion sample on its own Interest and Skill scales also is provided on the profile for the Occupational Scales. Finally each of the Orientation pages includes a short interpretive narrative that summarizes the individual's results.

The measurement of both interests and confidence in skills enriches the interpretive information that can be gleaned from the CISS scores. Based on a comparison of the level of the Interest and Skill scores for each scale, the individual is advised to *Pursue* the area if both the Interest and Skill scores are high, to *Develop* the area if the Interest score is high but the Skill score is low, to *Explore* the area if Interest is low and Skill high, or to *Avoid* if both Interest and Skill scores are low.

*Validity and reliability.* Substantial evidence of the construct validity of the Interest and Skill scales is presented in the manual of the CISS ([Campbell et al.,](#)

1992). Test-retest correlations over a 90-day interval are 0.87, 0.83, and 0.87 for the Orientation, Basic, and Occupational Interest Scales, respectively, and 0.81, 0.79, and 0.79 for the Orientation, Basic, and Occupational Skill Scales. Independently conducted studies report evidence of construct validity for both the Interest and Skill scale scores (Hansen & Leuty, 2007; Pendergrass, Hansen, Neuman, & Nutter, 2003; Sullivan & Hansen, 2004).

## Holland's interest inventories

The emergence of John Holland's theory of careers (Holland, 1959, 1966, 1973, 1997) began with the development of the Vocational Preference Inventory (VPI; Holland, 1958). Based on interest data collected with the VPI as well as data from other interest, personality, and values inventories and from analyses of the structure of interests, Holland formulated his theory of vocational life and personality. According to Holland, people can be divided into six types or some combination of six types: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. Holland indicates that the types can be organized in the shape of a hexagon in the R-I-A-S-E-C order; the types adjacent to one another on the hexagon (e.g., Realistic-Investigative or Enterprising-Conventional) are more related than types that are diametrically opposed to one another (e.g., Realistic-Social or Artistic-Conventional). Attempts to verify Holland's hexagonal representation of the world of work show in general that the structure of interests approximates the theoretical organization proposed by Holland (Campbell & Hansen, 1981; Hansen, Collins, Swanson, & Fouad, 1993; Haverkamp, Collins, & Hansen, 1994; Rounds, 1995).

Holland's theory has led to development of inventories and sets of scales to measure his six types, for example, his own SDS (Holland, 1971); the GOT of the SII (Campbell & Holland, 1972; Hansen & Johansson, 1972), the General Themes of the CAI (Johansson, 1975), and the Orientation Scales on the CISS (Campbell et al., 1992).

The VPI, developed by Holland (1958), is the predecessor to the SDS. The VPI was based on a series of theoretical and empirical reports including the personality, vocational choice, and vocational interest literatures. From the extant literature Holland was able to identify interest-personality factors and also to hypothesize how the types related to one another. Holland used the descriptions of the six types to develop a pool of 160 items that represented the interest factors or types. The result (Holland, 1985b) was seven homogeneous scales, constructed in a series of rational-empirical steps that measure Self-Control (SC) plus the six types hypothesized in Holland's theory: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C).

The SDS (Holland & Messer, 2013b; Holland, 1971, 1985a, 1994), similar to the VPI, was developed to measure Holland's six types. The SDS may be self-administered, self-scored, and to a limited degree self-interpreted. The 139-item assessment booklet includes four sections: Activities the respondent would like to do; Competencies, Occupations, and Self-Estimates.

The Fifth Edition of the SDS is suitable for those who are 11 years of age or older and have a seventh-grade reading level; Form Easy (E) is rated at the fourth-grade level. The SDS has been translated into more than 25 languages (Holland & Messer, 2013b). In addition to the popular paper-and-pencil administered and client scored version, the SDS also can be taken through the publisher's SDS-dedicated website ([www.Self-Directed-Search.com](http://www.Self-Directed-Search.com)) and through PARiConnect, an online assessment system. The most important feature of the SDS profile is the summary codes. The three highest raw scores represent the respondent's primary, secondary, and tertiary code assignments. Holland (1979) suggests flexibility in using the three summary codes for occupational exploration, since the codes are approximate, not precise.

A series of materials has been developed to assist in the interpretation of the SDS. The 2013 manuals (Holland & Messer, 2013a, 2013b), explain the use of the SDS for individual and group career assistance. *The Occupations Finder* (Holland & Messer, 2013c), and *The Educational Opportunities Finder* (Messer & Holland, 2013) provide three letter Holland codes for more than 1300 occupations and hundreds of junior, community, 4-year, and postgraduate majors, respectively. *The Veterans and Military Occupations Finder* (Messer, Greene, & Holland, 2013) provides Military Occupational Classifications (MOC) for the branches of the military; the MOC are coded with two-letter Holland types. Correspondence of military jobs with an occupation is generated using the Holland codes to link military and civilian jobs. Occupational and educational alternatives can be identified by surveying the two booklets to find possibilities with summary codes that are similar to the individual's summary codes. The *Leisure Activity Finder* (Messer, Greene, Kovacs, & Holland, 2013) links 840 leisure activities to Holland's six types.

*Reliability and validity.* The median test-retest reliability coefficient for the SDS scales over a 2-week interval is 0.95 and is 0.88 over 2- to 4-month periods (Holland & Messer, 2013b). Studies of the predictive validity of the SDS, for choice of occupation and college major over 1-, 2-, and 3-year intervals, indicate a range of accuracy from 35% to 66% (Holland & Messer, 2013b; Holland, 1979, 1985a).

## Strong Interest Inventory

The earliest version of the Strong Vocational Interest Blank (Strong, 1927) used the empirical method of contrast groups to construct Occupational Scales representing the interests of men in 10 occupations. The first form for women was published in 1933, and until 1974 the instrument was published with separate forms for women and men. In 1974 (Campbell, 1974), the two forms were combined by selecting the 325 best items from the previous women's (TW398) and men's (T399) forms, and in 1981 (Campbell & Hansen, 1981) another revision was completed in an effort to provide matched-sex Occupational Scales (e.g., male- and female-normed Forester

scales, male- and female-normed Flight Attendant scales, male- and female-normed Personnel Director scales). The 1985 revision ([Hansen & Campbell, 1985](#)) marked the end of the sex-equalization process which began in 1971. One additional major change in the 1985 revision was the expansion of the breadth of the profile to include more nonprofessional and vocational/technical occupational scales. The most recent revisions of the SII were completed in 2004 and 2012 ([Donnay et al., 2004](#); [Herk & Thompson, 2012](#)). The SII is available only through online assessment (the paper-and-pencil version was discontinued in 2015).

*Item pool and profile.* The item booklet for the current SII includes 291 items, divided into six sections. Part 1, Occupational Titles; Part 2, School Subjects; Part 3, Activities; Part 4, Leisure Activities; Part 5, Types of People, and Part 6, Personal Characteristics. The item format requires respondents to indicate the degree of their interest in each item by responding “Strongly Like,” “Like”, “Indifferent,” “Dislike,” and “Strongly Dislike.” The profile includes four sets of scales: six GOT, 30 Basic Interest Scales (BIS), 260 female and male Occupational Scales that represent 130 professional and nonprofessional occupations (e.g., farmers/ranchers, geographers, photographers, social workers, buyers, credit managers), and five Personal Styles Scales.

*Occupational Scales.* The Occupational Scales of the Strong are another example of test construction using the empirical method of contrast groups. The response rate percentage of the occupational criterion sample to each item is compared to the response rate percentage of the appropriate-sex contrast sample (i.e., General Reference Sample of females or males) to identify items that differentiate the two samples. For the 2004 and 2012 SII, 12–45 items were identified as the interests (“Likes”) or the aversions (“Dislikes”) of each occupational criterion sample. The raw scores for an individual scored on the Occupational Scales are converted to standard scores based on the occupational criterion sample, with mean set equal to 50 and standard deviation of 10. For all occupations, matched-sex scales are presented on the SII profile.

*General Occupational Themes.* The GOT are a merger of Strong’s empiricism with Holland’s theory of vocational types. The six homogeneous Themes contain items selected to represent Holland’s definition of each type—Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. Data comparing the enhanced 2004 and 2012 GOT to Holland’s VPI or SDS are not available. However, the 1974 GOT correlated highly (0.72–0.79) with same-named VPI scales ([Hansen, 1983](#)). Correlations between the GOT indicate that the hexagonal order that Holland proposed to describe the relationship between his types (adjacent types have more in common than do diametrically opposed types) also describes the relationships among the SII Themes ([Harmon, Hansen, Borgen, & Hammer, 1994](#)).

The score information provided for the GOT includes the standard scores, based on a General Reference Sample composed of women and men with mean set equal to 50 and standard deviation of 10. In addition to standard scores, interpretive information on the report provides a visual representation of the distribution of the female General Reference Sample and male General Reference Sample on each GOT.

The integration of Holland’s theory with Strong’s empiricism provides the organizational framework for the current Strong profile. The Occupational Scales are

coded with one to three Holland types based on the criterion sample's highest scores on the GOT. The codes, in turn, are used to categorize the Occupational Scales on the profile. The BIS also are clustered according to Holland types by identifying the Theme with which each BIS has its highest correlation.

*Basic Interest Scales.* The original 25 BIS were constructed using the statistical technique of cluster analysis to identify highly correlated items (Campbell, Borgen, Eastes, Johansson, & Peterson, 1968). The BIS were developed to focus on the measurement of only one interest factor per scale, and, consequently, are easier to interpret than the heterogeneous Occupational Scales that incorporate items, representing several interest factors as well as likes and aversions, in each scale.

The scale names for the current 30 BIS describe the homogeneous item content and the interest trait measured by each scale. Like the GOT, standard scores based on a combined-sex General Reference Sample and interpretive information based on female and male General Reference Samples is presented in the report.

*Personal Style Scales.* Five Personal Styles Scales—Team Orientation, Work Style, Learning Environment, Leadership, and Risk-Taking—also are reported on the Strong profile. All five of these bipolar scales were standardized using the combined-sex General Reference Sample; interpretive information based on female and male General Reference Samples are presented in the SII report.

The Team Orientation scale taps a person's interest in working independently and identifies those who have a preference for self-reliance (low scores) as opposed to those who prefer working with others and collaborating (high scores). The Work Style Scale is intended to identify people who prefer to work with ideas, data, and things (low scores) and those who prefer to work with people (high scores). The Learning Environment Scale distinguishes between people who prefer academic learning environments (high scores) and those who prefer practical training (low scores). Similarly, the Leadership Scale is meant to identify those who prefer to do a task themselves or to lead by example (low scores) and those who like to be in charge of others (high scores). The Risk-Taking Scale, as the scale name implies, measures the extent to which an individual is willing to take risks.

*Reliability and validity.* The test–retest reliability of the scale scores on the SII scales is substantial. Median reliabilities over short and long interval periods for the GOT were 0.86, and 0.84; for the BIS were 0.74–0.94; and for the Occupational Scales the range was from 0.71 to 0.93 (Donnay et al., 2004).

Because interest inventories are used to make long-term decisions, predictive validity is important. The SII has a long history of predictive validity studies for its various editions; however, no predictive validity data are available at this time for the 1994, 2004, or 2012 editions of the SII. Data from earlier forms of the Strong show that, at least in the past, high scores on the Occupational Scales are related to occupations eventually entered; generally, between one-half and three-fourths of the subjects in predictive validity studies enter occupations predictable from their earlier scores (Campbell, 1966; Hansen & Dik, 2005; Spokane, 1979). Studies assessing the usefulness of the SII for predicting college majors have found hit rates similar to those reported for occupational entry (Hansen & Lee, 2007; Hansen & Swanson, 1983; Hansen & Tan, 1992).

## Career Assessment Inventory

The first edition of the CAI ([Johansson & Johansson, 1978](#); [Johansson, 1975](#)) was developed for use with individuals considering immediate career entry, community college education, or vocational-technical training, and was modeled after the SII. In 1982 the decision was made to move from separate-sex to combined-sex Occupational Scales. The enhanced version of the CAI published in 1986 ([Johansson, 1986](#)) was expanded to include several Occupational Scales representing professional occupations.

The enhanced CAI test booklet includes 370 items, and the profile reports three sets of scales: six homogeneous General Themes, 25 homogeneous Basic Interest Areas and 111 heterogeneous Occupational Scales. The CAI uses Holland's theory to organize the Basic Interest Areas and Occupational Scales on the profile, clustering together those that represent each of Holland's six types.

The Enhanced Version of the CAI also has an interpretive report option that provides interpretive text in addition to the standard profile. The interpretive report provides suggestions for linking Occupational Scale scores to occupations that do not appear on the profile and provides referrals to career-search resources on the Internet.

The CAI has three scoring options. The traditional paper-and-pencil version requires sending the completed answer sheet to the publisher for scoring. The profile is returned by US Mail. Desktop software from the publisher can be used to score assessments, report results, and store data on the provider's computer. The Web-based administration provides quick assessment and reports online.

*CAI Scales.* The Occupational Scales of the enhanced version of the CAI were developed using the empirical method of contrast samples to select items that differentiated combined-sex criterion and general reference samples from each other. The combined-sex criterion samples fall short of the goal of equally representing females and males within the sample (e.g., 0 female aircraft mechanics, 0 male medical assistants, 0 female purchasing agents, and 0 male secretaries). The author attempted to improve the psychometrics of the scales by doing separate-sex item analyses if the separate-sex samples were large enough. In most instances, however, the sample representing one sex or the other was too small (e.g., 22 male bank tellers, 16 female dental laboratory technicians, 12 male bookkeepers, and 20 female enlisted personnel) to produce reliable item analyses. In fact, 64 of the 111 Occupational Scales were developed using criterion samples that included less than 50 subjects representing one sex or the other (45 scales with <50 female subjects and 19 scales with <50 male subjects). Consequently, the exploration validity for the scales developed with unbalanced female–male ratio criterion samples, is questionable for the underrepresented gender.

The General Themes and Basic Interest Areas are normed on a combined-sex reference sample composed of employed adults and students drawn from the six Holland interest areas: 75 females and 75 males from each of the six types, for a total of 900 subjects, compose the sample. In addition to the standard scores based on a combined-sex sample, the CAI profile presents bars for each scale representing the range of scores for females and males in the reference sample. These additional

data help to circumvent the problem of gender differences on some of the homogeneous scales.

The CAI also includes four nonoccupational scales. The first of these, the Fine Arts-Mechanical scale identifies people interested in skilled trade occupations (high scores) and creative and social service occupations (low scores). The Occupational Extroversion–Introversion Scale differentiates people who prefer to work alone (high scores) and those who like working with people (low scores). The Educational Orientation Scale identifies people whose interests are associated with those of students in higher education (high scores) and those who choose on-the-job training (low scores). The Variability of Interests Scale reflects widespread and diverse interests (high scores) or more narrow, focused interests (low scores).

*Reliability and validity.* The evidence of test–retest reliabilities of scores on the CAI resemble those found for other interest inventories. The median test–retest for the six General Themes is 0.93 over a 2-week interval and 0.81 over 6–7 years. For the Basic Interest Areas the median test–retest over 2 weeks is 0.90 and over 6–7 years is 0.77, and for the Occupational Scales the median test–retest is 0.92 and 0.82, respectively, for the same intervals (Johansson, 1984). Evidence of validity for the CAI scales is primarily convergent in nature and shows evidence of robust correlations, among the Basic Interest Areas and Occupational Scales and the General Themes, that are similar to those found with the SII and the CISS.

## O\*NET Interest Profiler and Interest Item Pool

In the 1990s, the US Department of Labor’s Office of Policy and Research replaced the decades old *Dictionary of Occupational Titles (DOT)* with the Occupational Information Network commonly known now as the O\*NET. The O\*NET includes numerous components designed to provide online access to career exploration.

The comprehensive O\*NET system organizes and describes data on the characteristics of occupations and the characteristics or attributes of workers. Some of the products available through the online O\*NET Resource Center include the My Next Move and My Next Move for Veterans, which are web applications for students, new job seekers, and veterans entering the civilian workforce, that provide career options to match interests and experiences. O\*NET Career Exploration Tools include The Ability Profile, The Work Importance Locator, and The Interest Profiler.

The Interest Profiler was developed first as a paper-and-pencil self-scoring vocational interest measure (Lewis & Rivkin, 1999). This initial 180 item version was followed by a computerized version (Rounds et al., 1999). All versions of The Interest Profiler report six scores for scales that measure Holland’s six RIASEC types. A short form of the O\*NET Interest Profile, that can be administered in either a paper-and-pencil or computerized form and an O\*NET Mini Interest Profiler (Mini-IP) for Mobile Devices also have been developed (Rounds, Ming, Cao, Song, & Lewis, 2016; Rounds, Su, Lewis, & Rivkin, 2010). The short form of the O\*NET Interest Profiler has 60 items, and the Mini-IP has only 30 items. The

National Center for O\*NET Development and the US Department of Labor's office of Policy and Research collaborated with researchers at the University of Illinois at Urbana-Champaign to create the shortened scales (Rounds et al., 2016). The Short Form of the Interest Profiler is linked to all the data on the O\*NET including the Occupational Interest Profiles (OIP; Rounds, Smith, Hubert, Lewis, & Rivkin, 1999). The OIPs incorporate Holland's six types to describe 1172 Occupational Units included in the O\*NET, and Holland's types provide the connection between the client's interests and the job requirements.

Akin to Goldberg and his colleagues (2006) development of a public domain item pool for personality assessment, Rounds at the University of Illinois worked with colleagues at several universities to develop The IIP. The items were written by graduate students working with Rounds and Armstrong (at Iowa State University). Through a series of analyses, two sets of scales were constructed to measure Holland's six types (Armstrong, Allison, & Rounds, 2008), and a set of 31 Basic Interest Markers scale also was constructed (Liao, Armstrong, & Rounds, 2008).

## ***Stability of interests***

The degree to which interests are stable is important to the predictive power of inventories. If interests are fickle and unstable, interest inventory scores will not explain any of the prediction variance.

Thus stability of interests was one of the earliest concerns of researchers in interest measurement (Strong, 1943). Cross-sectional and longitudinal methods have been used in a plethora of studies to document that interests are stable even at relatively young ages of 15 or 16 years. By age 20, the stability of interests is obvious even over test-retest intervals of 5–10 years, and by age 25, interests are very stable (Hansen & Swanson, 1983; Hansen, 2013a; Johansson & Campbell, 1971; Swanson & Hansen, 1986). During the long history of the SII, over 30 occupations have been tested at least three times: in the 1930s, the 1960s, and the 1970s/80s. Analyses of these data have shown that interests of randomly sampled occupational groups are stable (Hansen, 1988). Fig. 6.2, a profile of interests for lawyers collected in the 1930s, 1960s, and 1970s illustrates the typical finding for all the occupations:

1. The configuration of the interests of an occupation stays the same over long periods of time.
2. Even when interests change to some small extent, the relative importance of various interests stays the same (Hansen, 1988).

Hansen and Leuty (2014) used three decades of archival data to explore the relations between age, birth cohort, and vocational interests. Contrary to the popular literature that often paints large differences between generations, they found that the overall amount of variance explained by birth year was less than 1% of the variance for any interest area. The interaction of age and birth cohort also explained less than 1% of the variance.

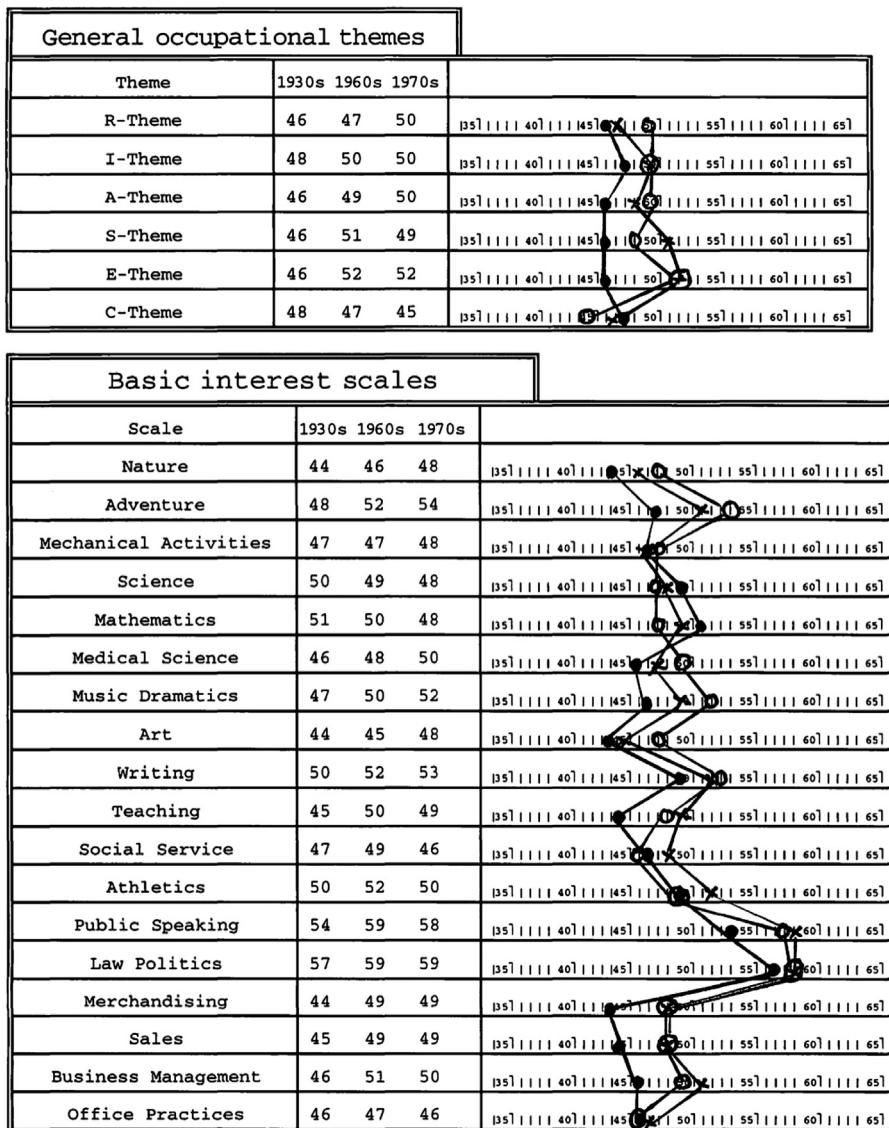


Figure 6.2 Mean interest profile for lawyers in the 1930s (\_\_\_\_\_), the 1960s (x\_\_\_\_x), and the 1970s (o\_\_\_\_o).

### Use of interest inventories

Interest inventories are used to efficiently assess interests by a variety of institutions including high school and college advising offices, social service agencies, employment agencies, consulting firms, corporations, and community organizations such as the YMCA.

## Career exploration

The major use of assessed interests, usually reported as interest inventory scores, is in career counseling that leads to decisions such as choosing a major, selecting an occupation, making a mid-career change, or preparing for retirement. First, counselors use the interest inventory reports to develop hypotheses about clients that may be discussed, confirmed, or discarded during career exploration. Then, the interest scores and profile provide a framework for interest exploration and a mechanism for helping the client to integrate her or his past history with current interests.

The inventory results serve as a starting point for evaluating interests, as an efficient method for objectively identifying interests, and as a structure for the counseling process. Inventory results help some counselees to increase the number of options they are considering; some use the results to begin to narrow the range of possible choices. Others only want to confirm educational or vocational decisions that they already have made. Gottfredson (1986) noted that interest inventory results also can be used to help clients explore and understand the way in which their aspirations, reflected in interest scores, are constrained by barriers and impediments such as poverty, lack of educational opportunities, family responsibilities, or physical and mental disabilities.

## Selection and placement

Interest inventories also are used to assess interests during employment selection and placement evaluations. Among qualified candidates, interest inventories help to identify those most likely to complete the training program and stay in the profession (Nye, Su, Rounds, & Drasgow, 2012; Van Iddekinge, Roth, Putka, & Laniwch, 2011). Even after initial selection, interest inventories may be used to help an employee find the right job within the company (Hansen, 2013b, 1994). Interests, along with personality, are receiving renewed attention from industrial–organizational psychologists (Hansen, 2013b, 2013c). In particular, organizations use interest assessment for career coaching, human resources management, and for mid-career change and retirement planning (Hansen & Wiernik, 2018).

## Research

Researchers use measures of interests (e.g., checklists, self-estimates, ratings scales, interest inventories) to operationalize interest traits, investigate the origin and development of interests, explore changes or stability in society, and understand the relationship between interests and other psychological variables such as abilities, satisfaction, success, and personality. Studies assessing the structure of interests and also the interests of various occupational groups provide information for understanding the organization of the world of work and the relationships among occupations.

Most interest inventories are constructed to measure vocational interests. Recent research, however, indicates that instruments such as the SII measure not only vocational interests but also leisure interests (Cairo, 1979; Varca & Shaffer, 1982). Holland (1973) has proposed that instruments measuring his six personality types also can identify a respondent's preferences for environments and types of people as well as job activities. The LIQ (Hansen, 1998) is one measure that was developed specifically to assess leisure interests. Beginning with a pool of over 700 items, a series of cluster analyses was used to reduce the item pool. Then, factor analysis was used to identify items for each of 20 homogeneous scales. The LIQ scales are analogous to the BIS on the SII, the Basic Scales on the CISS, and the Basic Area Scales on the CAI. The median internal consistency for the scales is 0.86 (range 0.93 for *Team Sports* to 0.69 for *Travel*) and the median test-retest over 5 weeks is 0.85 (range 0.91 for *Hunting & Fishing* and *Cultural Arts* to 0.61 for *Travel*). The LIQ scales correlate as expected with the BIS and GOT of the SII. For example, *Team Sports* on the LIQ correlates 0.79 with the SII *Athletics* BIS, and LIQ *Building and Restoring* correlates 0.77 with the *Realistic* GOT (Hansen & Scullard, 2002).

### **Future directions**

The frequency of test use in counseling has not changed appreciably in the last 40 years. A wide variety of new interpretive materials, career guidance packages, and interactive computerized systems for inventory interpretation and career exploration are available. Thus far, evaluations of the use of interest inventories indicate that various modes and mediums of presentation are equally effective (Hansen, Neuman, Haverkamp, & Lubinski, 1997; Vansickle & Kapes, 1993). The trend in the future, with decreasing budgets and personnel in educational institutions, will be toward even greater use of computers for interest inventory administration and interpretation and for integration into computerized career counseling modules.

Techniques for developing reliable and valid interest inventories are available now, and the construction methods have reached a plateau of excellence in reliability and validity. Therefore publishers can direct their efforts toward an increased emphasis on interpretation and counselor competency. Test manuals traditionally were written to provide data required by the American Psychological Association's "Standards for Educational and Psychological Testing" (2014); now, interpretive manuals are prepared in addition to technical manuals to help the professional maximize the usefulness of inventory results (Hansen, 1992; Holland & Messer, 2013a). Increasingly publishers are attempting to develop testing packages that integrate interest inventories or measures of other constructs (e.g., the SII and the Myers-Briggs Type Indicator). Unfortunately, these packages have been released by publishers without expending much effort to collect data to assess the validity of using the instruments as a package.

Job design is a promising area for research that incorporates the use of interest inventories. This approach would focus on identifying the interactions between interests and workplace characteristics. Similarly, understanding the interaction of

interests with approaches to workplace training may help educational and business organizations develop more effective training methods (Hansen & Wiernik, 2018).

As the use of interest inventories expands to new populations, research also must move in that direction to aid in understanding the characteristics of the populations as well as the best methods for implementing interests inventories with them. The cross-cultural use of interest inventories also is increasing the demand for valid translations of inventories and for data on the predictive accuracy of inventories normed on US populations for non-English-speaking respondents (Fouad & Spreda, 1995; Hansen 2013b).

## Summary

Interest inventories will be used in the future as in the past to operationalize the trait of interests in research. Attempts to answer old questions, such as the interaction of interests and personality, success, values, satisfaction, and ability will continue. Holland's theory undoubtedly will continue to evoke research in the field. Studies designed to understand educational and vocational dropouts and changers, to analyze job satisfaction, to understand the development of interests, and to predict job or academic success will draw on Holland's theoretical constructs for independent variables and on interest inventories to identify interests. Exploration of vocational interests always has been a popular topic in counseling psychology. However, the cost of using commercial interest inventories, even with research discounts, often exceeds the resources of academic researchers who in the past have been the major contributors to understanding both the construct of interests as well as the measurement of interests. As a result, research on interests has waned considerably over the past 20 years. The development of the IIP has the potential to renew research on interests in the same way that the IPIP has helped to stimulate personality research.

## References

- American Psychological Association. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Armstrong, P. I., Allison, W., & Rounds, J. B. (2008). Development and initial validation of brief public domain RIASEC marker scales. *Journal of Vocational Behavior*, 73, 287–299.
- Cairo, P. C. (1979). The validity of the Holland and basic interest scales of the Strong vocational interest blank: Leisure activities versus occupational membership as criteria. *Journal of Vocational Behavior*, 15, 68–77.
- Campbell, D. P. (1966). Occupations ten years later of high school seniors with high scores on the SVIB life insurance salesman scale. *Journal of Applied Psychology*, 50, 369–372.
- Campbell, D. P. (1974). *Manual for the SVIB–SCII*. Stanford, CA: Stanford University Press.

- Campbell, D. P. (1995). The Campbell interest and skill survey (CISS): A product of ninety years of psychometric evolution. *Journal of Career Assessment*, 3, 391–410.
- Campbell, D. P., Borgen, F. H., Eastes, S. H., Johansson, C. B., & Peterson, R. A. (1968). A set of basic interest scales for the strong vocational interest blank for men. *Journal of Applied Psychology Monograph*, 52, 1–54.
- Campbell, D. P., & Hansen, J. C. (1981). *Manual for the SVIB–SCII* (3rd ed.). Stanford, CA: Stanford University Press.
- Campbell, D. P., & Holland, J. L. (1972). Applying Holland's theory to Strong's data. *Journal of Vocational Behavior*, 2, 353–376.
- Campbell, D. P., Hyne, S. A., & Nilsen, D. L. (1992). *Manual for the Campbell interest and skill survey*. Minneapolis, MN: National Computer Systems.
- Consulting Psychologists Press. (2004). *Strong interest inventory*. Palo Alto, CA: CPP.
- Consulting Psychologists Press. (2012). *Strong interest inventory*. Palo Alto, CA: CPP.
- Cowdery, K. M. (1926). Measurement of professional attitudes: Differences between lawyers, physicians, and engineers. *Journal of Personnel Research*, 5, 131–141.
- Dik, B. J., & Hansen, J. C. (2008). Following passionate interests to well-being. *Journal of Career Assessment*, 16, 86–100.
- Donnay, D., Schaumbhut, N., Thompson, R., Harmon, L., Hansen, J. C., Bergen, F., & Hammer, A. (2004). *Strong interest inventory manual*. Palo Alto, CA: Consulting Psychologists Press.
- Fouad, N. A., & Spreda, S. L. (1995). Use of interest inventories with special populations: Women and minority groups. *Journal of Career Assessment*, 4, 453–468.
- Freyd, M. (1922–1923). The measurement of interests in vocational selection. *Journal of Personnel Research*, 1, 319–328.
- Gottfredson, L. S. (1986). Special groups and the beneficial use of vocational interest inventories. In W. B. Walsh, & S. H. Osipow (Eds.), *Advances in vocational psychology*. Hillsdale, NJ: Lawrence Erlbaum.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. E. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40, 84–96.
- Guilford, J. P., Christensen, P. R., Bond, N. A., Jr., & Sutton, M. A. (1954). A factor analysis study of human interests. *Psychological Monographs*, 375(68), 1–38.
- Hansen, J. C. (1988). Changing interests: Myth or reality? *Applied Psychology: An International Review*, 37, 133–150.
- Hansen, J. C. (1992). *User's guide to the SII* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.
- Hansen, J. C. (1983). Correlation between VPI and SCII scores. Unpublished manuscript, Center for Interest Measurement Resarch, Univeristy of Minnesota.
- Hansen, J. C. (1994). The measurement of vocational interests. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 293–316). Hillsdale, NJ: Lawrence Erlbaum.
- Hansen, J. C. (1998). *Leisure interest questionnaire*. St. Paul, MN: JCH Consulting.
- Hansen, J. C. (2013a). Nature, importance, and assessment of interests. In S. Brown, & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (2nd ed., pp. 281–304). Hoboken, NJ: Wiley & Sons, Inc.
- Hansen, J. C. (2013b). Person–environment fit approach to cultivating meaning. In B. J. Dik, Z. S. Byrne, & M. F. Steger (Eds.), *Purpose and meaning in the workplace* (pp. 37–55). Washington, DC: American Psychological Association.

- Hansen, J. C. (2013c). Personality and vocational behavior. In R. Tett, & N. Christiansen (Eds.), *Handbook of personality at work* (pp. 651–670). New York, NY: Routledge.
- Hansen, J. C. (2016). Career counseling with adults: Theories, interventions and populations. In S. Maltzman (Ed.), *The Oxford handbook of treatment processes and outcomes in psychology* (pp. 372–403). New York: Oxford University Press.
- Hansen, J. C., & Campbell, D. P. (1985). *Manual for the SVIB–SCII* (4th ed.). Stanford, CA: Stanford University Press.
- Hansen, J. C., Collins, R., Swanson, J. L., & Fouad, N. A. (1993). Gender differences in the structure of interests. *Journal of Vocational Behavior*, 42, 200–211.
- Hansen, J. C., & Dik, B. (2005). Evidence of 12-year predictive and concurrent validity for SII occupational scale scores. *Journal of Vocational Behavior*, 67, 365–378.
- Hansen, J. C., & Johansson, C. B. (1972). The application of Holland's vocational model to the strong vocational interest blank for women. *Journal of Vocational behavior*, 2, 479–493.
- Hansen, J. C., & Lee, W. V. (2007). Evidence of concurrent validity of SII scores for Asian American college students. *Journal of Career Assessment*, 15, 1–11.
- Hansen, J. C., & Leuty, M. E. (2007). Evidence of validity for the skill scales of the Campbell interest and skill survey. *Journal of Vocational Behavior*, 71, 23–44.
- Hansen, J. C., & Leuty, M. E. (2014). Teasing apart the relations between age, birth cohort, and vocational interests. *Journal of Counseling Psychology*, 61, 289–298.
- Hansen, J. C., Neuman, J., Haverkamp, B. E., & Lubinski, B. R. (1997). Comparison of user reaction to two methods of SII administration and report feedback. *Measurement and Evaluation in Counseling and Development*, 30, 115–127.
- Hansen, J. C., & Scullard, M. G. (2002). Psychometric evidence for the Leisure interest questionnaire and an analyses of the structure of leisure interests. *Journal of Counseling Psychology*, 49, 331–341.
- Hansen, J. C., & Swanson, J. L. (1983). Stability of interests and the predictive and concurrent validity of the 1981 Strong-Campbell interest inventory. *Journal of Counseling Psychology*, 30, 194–201.
- Hansen, J. C., & Tan, R. N. (1992). Concurrent validity of the 1985 strong interest inventory for college major selection. *Measurement and Evaluation in Counseling and Development*, 25, 53–57.
- Hansen, J. C., & Wiernik, B. (2018). Work preferences: Vocational interests and values. In (2nd ed.) N. Anderson, D. S. Ones, H. K. Sinahgil, & C. Viswesvaran (Eds.), *The Sage handbook of industrial, work and organizational psychology: Personnel psychology and employee performance* (Vol. 1). Thousand Oaks, CA: Sage Publications.
- Harmon, L., Hansen, J. C., Borgen, F., & Hammer, A. (1994). *Strong interest inventory applications and technical guide*. Stanford, CA: Stanford University Press.
- Haverkamp, B. E., Collins, R. C., & Hansen, J. C. (1994). Structure of interests of Asian-American college students. *Journal of Counseling Psychology*, 41, 255–264.
- Herke, N. A., & Thompson, R. C. (2012). *Strong interest inventory manual supplement*. Mountain View, CA: Consulting Psychologists Press.
- Holland, J. L. (1958). A personality inventory employing occupational titles. *Journal of Applied Psychology*, 42, 36–342.
- Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology*, 6, 35–45.
- Holland, J. L. (1966). *The psychology of vocational choice*. Waltham, MA: Blaisdell.
- Holland, J. L. (1971). *The counselor's guide to the self-directed search*. Palo Alto, CA: Consulting Psychologists Press.

- Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Englewood Cliffs, NJ: Prentice-Hall.
- Holland, J. L. (1979). *The self-directed search professional manual*. Palo Alto, CA: Consulting Psychologists Press.
- Holland, J. L. (1985a). *Professional manual for the self-directed search*. Odessa, FL: Psychological Assessment Resources.
- Holland, J. L. (1985b). *Vocational preference inventory (VPI) manual—1985 edition*. Odessa, FL: Psychological Assessment Resources.
- Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.
- Holland, J. L. (1994). *Self-directed search form R: 1994 edition*. Odessa, FL: Psychological Assessment Resources.
- Holland, J. L., & Messer, M. A. (2013a). *John Holland's self-directed search fast guide*. Lutz, FL: Psychological Assessment Resources, Inc.
- Holland, J. L., & Messer, M. A. (2013b). *John Holland's self-directed search professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.
- Holland, J. L., & Messer, M. A. (2013c). *The occupations finder*. Lutz, FL: Psychological Assessment Resources, Inc.
- Hubbard, R. M. (1930). Interest analysis blank. In D. G. Paterson, R. M. Elliott, L. D. Anderson, H. A. Toops, & E. Heidbrider (Eds.), *Minnesota mechanical ability test*. Minneapolis, MN: University of Minnesota Press.
- Johansson, C. B. (1975). *Manual of the career assessment inventory*. Minneapolis, MN: National Computer Systems.
- Johansson, C. B. (1984). *Manual for career assessment inventory* (2nd ed.). Minneapolis, MN: National Computer Systems.
- Johansson, C. B. (1986). *Career assessment inventory: Enhanced version*. Minneapolis, MN: National Computer Systems.
- Johansson, C. B., & Campbell, D. P. (1971). Stability of the strong vocational interest blank for men. *Journal of Applied Psychology*, 55, 24–26.
- Johansson, C. B., & Johansson, J. C. (1978). *Manual supplement for the career assessment inventory*. Minneapolis, MN: National Computer Systems.
- Kitson, H. D. (1925). *The psychology of vocational adjustment*. Philadelphia, PA: Lippincott.
- Kornhauser, A. W. (1927). Results from a quantitative questionnaire of likes and dislikes used with a group of college freshmen. *Journal of Applied Psychology*, 11, 85–94.
- Lewis, P., & Rivkin, D. (1999). *O\*NET interest profiler*. Raleigh, NC: National Center for O\*NET Development.
- Liao, H.-Y., Armstrong, P. I., & Rounds, J. B. (2008). Development and initial validation of public domain basic interest markers. *Journal of Vocational Behavior*, 73, 159–183.
- Lunneborg, P. W. (1976). *Manual for the vocational interest survey*. Seattle, WA: University of Washington, Educational Assessment Center.
- Messer, M. A., Greene, J. A., & Holland, J. L. (2013). *Veterans and military occupations finder*. Lutz, FL: Psychological Assessment Resources, Inc.
- Messer, M. A., Greene, J. A., Kovacs, A. M., & Holland, J. L. (2013). *The leisure activities finder*. Lutz, FL: Psychological Assessment Resources, Inc.
- Messer, M. A., & Holland, J. L. (2013). *The educational opportunities finder*. Lutz, FL: Psychological Assessment Resources, Inc.
- Miner, J. B. (1922). An aid to the analysis of vocational interests. *Journal of Educational Research*, 5, 311–323.

- Nye, C. D., Su, R., Rounds, J. B., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives in Psychological Science*, 7, 384–403.
- Parsons, F. (1909). *Choosing a vocation*. Boston, MA: Houghton Mifflin.
- Paterson, D. G., Elliott, R. M., Anderson, L. D., Toops, H. A., & Heidbreder, E. (Eds.). (1930). Minnesota mechanical abilities test. Minneapolis, MN: University of Minnesota Press.
- Pendergrass, L., Hansen, J. C., Neuman, J., & Nutter, K. (2003). Examination of the concurrent validity of scores from the CISS for student-athlete college major selections: A brief report. *Measurement and Evaluation in Counseling and Development*, 35, 212–217.
- Remmers, H. H. (1929). The measurement of interest differences between students of engineering and agriculture. *Journal of Applied Psychology*, 13, 105–119.
- Roe, A. (1956). *The psychology of occupations*. New York, NY: Wiley.
- Rounds, J. B. (1995). Vocational interests. Evaluating structural hypotheses. In R. V. Dawis, & D. Lubinski (Eds.), *Assessing individual differences in human behavior* (pp. 177–232). Palo Alto, CA: Consulting Psychologists Press.
- Rounds, J. B., Mazzeo, S. E., Smith, T. J., Hubert, L., Lewis, P., & Rivkin, D. (1999). *O\*NET interest profiler: Reliability, validity, and self-scoring*. Raleigh, NC: National Center for O\*NET Development.
- Rounds, J. B., Ming, C. W. J., Cao, M., Song, C., & Lewis, P. (2016). *Development of an O\*NET mini interest profile (Mini-IP) for mobile devices: Psychometric characteristics*. Raleigh, NC: National Center for O\*NET Development.
- Rounds, J. B., Smith, T., Hubert, L., Lewis, P., & Rivkin, D. (1999). *Development of occupational interest profiles for O\*NET*. Raleigh, NC: National Center for O\*NET Development.
- Rounds, J. B., Su, R., Lewis, P., & Rivkin, D. (2010). *O\*NET interest profiler short form psychometric characteristics*. Raleigh, NC: National Center for O\*NET Development.
- Spokane, A. R. (1979). Occupational preferences and the validity of the Strong-Campbell Interest Inventory for college women and men. *Journal of Counseling Psychology*, 26, 312–318.
- Strong, E. K., Jr. (1927). *Vocational interest blank*. Stanford, CA: Stanford University Press.
- Strong, E. K., Jr. (1943). *Vocational interests of men and women*. Stanford, CA: Stanford University Press.
- Sullivan, B., & Hansen, J. C. (2004). Evidence of construct validity of the interest scales on the Campbell interest and skill survey. *Journal of Vocational Behavior*, 65, 179–202.
- Swanson, J. L., & Hansen, J. C. (1986). A clarification of Holland's construct of differentiation: The importance of score elevation. *Journal of Vocational Behavior*, 28, 163–173.
- Ton, M. N., & Hansen, J. C. (2001). Using a person–environment fit framework to predict satisfaction and motivation in work and marital roles. *Journal of Career Assessment*, 9, 315–331.
- Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology*, 96, 1167–1194.
- Vansickle, T. R., & Kapes, J. T. (1993). Comparing paper–pencil and computer-based versions of the Strong-Campbell interest inventory. *Computers in Human Behavior*, 9, 441–449.
- Varca, P. E., & Shaffer, G. S. (1982). Holland's theory: Stability of avocational interests. *Journal of Vocational Behavior*, 21, 288–298.

## **Part V**

# **Neuropsychological Assessment**

# Sources of error and meaning in the pediatric neuropsychological evaluation\*

7

Michael D. Weiler<sup>1</sup>, W. Grant Willis<sup>2</sup> and Mary Lynne Kennedy<sup>3</sup>

<sup>1</sup>Cranston Public School System, Cranston, RI, United States, <sup>2</sup>University of Rhode Island, Kingston, RI, United States, <sup>3</sup>Neuropsychology Partners Incorporated, East Providence, RI, United States

*“One of the main features of human life is a permanent development.”*

*Anokhin (1968, p. 81)*

Many child neuropsychological evaluations begin at the end, that is, they begin with the question they need to be able to answer when the evaluation is completed. At times, the question is “What are the real-world difficulties that we expect this child to have as a consequence of a particular medical condition (e.g., brain tumor, traumatic brain injury, lead exposure, etc.)?” At other times, the question falls more in the realm of “What explains the pattern of difficulties this child is having in his or her life?” Therefore, when we talk about neuropsychology being about brain–behavior relationships, we are interested in the functioning of the child and his or her brain operating in the real world. The child neuropsychological evaluation attempts to create less complicated analogues of these real-life behaviors so that performances on isolated functions can be compared against performances of other children the same age and, in this way, dissect the much more complicated real-world behaviors into their component parts. When we understand the ways these components fit together we should be able to explain not only why the child is having difficulty with this set of real-world behaviors, but also why this child is able to accomplish what he or she has accomplished despite these weaknesses.

The assessment of the child has to take into account what Jane Bernstein ([Bernstein & Weiler, 2000](#)) has called the cardinal feature of the child, which is development. Whereas with adults, neuropsychological evaluations are often requested for someone who has lost functions, in the majority of nonmedical referrals, the evaluation is requested on a youngster who, for some reason, is off developmental track and has not acquired milestones (social, academic, real-world functioning) that we would have expected him or her to have developed. When these areas of deficit are identified, the critical question is whether this is an

\*This chapter is dedicated to the memory of Michael D. Weiler, a beloved husband and father and esteemed colleague, who died shortly after its completion.

instance of the child being off developmental track and likely to catch up, or if this represents a more persistent deficit. In order to fully understand the phenomenon of this child's experience, we need to delve even further and consider the interaction between the child's real-life experience and its effect on the developing brain. It is for these reasons that the child neuropsychological evaluation needs to be so broad in scope: not only to deconstruct and assess the range of real-world functions that are relevant to the referral question, but also to understand the child and his or her developmental and experiential history, whether there are differences in the way that his or her behavior manifests across settings, and the degree to which effort, rapport, emotional, and motivational factors influence the findings.

As a practical matter, knowledge about the child in his or her environment is problematic yet this is a critical piece of information for the assessment. The number of multiple observations required to obtain a stable sample in even one setting is large. As a result, we are required to use questionnaires and reports of the adults around the child, but these have their own issues with consistency and observer perspective as evidenced by the low inter-rater reliability among informants. The contradiction here is that although a large amount of information is required to develop a broad understanding of the child, our human information-processing capabilities are not equipped to integrate and weigh so many sources of information simultaneously. We invariably use shortcuts and heuristics, but these rules of thumb, which typically are helpful in managing our behavior in the real world, do not always work as well for the complex task of making sense of the child's developing brain or answering the question of why the child is having a particular difficulty.

This chapter is not intended to provide the reader with a comprehensive, step-by-step manual on how to conduct a child neuropsychological evaluation. Rather, we address what we view as important yet sometimes underappreciated and under-considered aspects of this process. We begin with a discussion of some historical foundations, then outline the purpose of the child neuropsychological evaluation, including some ways in which it differs from the more common psychoeducational evaluation typically conducted in school systems. We continue with a discussion of the process of the evaluation, and conclude with a discussion of sources of error that may obfuscate the validity of child neuropsychological evaluations. Here, issues of incremental validity, demographic characteristics of children, ecological validity, malingering, and errors associated with clinical decision making are summarized.

## **Historical foundation**

Anokhin (1968, p. 81, cited in [Glozman, 2013](#)) observed that “one of the main features of human life is a permanent development.” Although ontogenetic development is characteristic of all phases of human life, perhaps we are more aware of it among children because it is more obvious from both a perspective of qualitative change and the speed with which changes occur. Before discussing some current

issues in child neuropsychology, however, it may be worthwhile to discuss some of the origins of the current strains of child neuropsychology (more western in its origin) and developmental neuropsychology with its Russian roots and Lev Semionovich Vygotsky, who we believe is a founding parent of both strains of neuropsychology.

### ***Origins of child neuropsychology***

Vygotsky is widely considered to be one of the most influential psychologists of the 20th century (Haagbloom et al., 2002). He was a mentor of and collaborator with Alexander R. Luria whose writings have had a profound influence in the field of adult neuropsychology (see also Benton & Sivan, 2007; Cubelli, 2005; Reitan, 1989 for other views). In contrast, Vygotsky's writings are more closely associated with learning, educational psychology, and developmental and child neuropsychology. If, as Bernstein has argued (Bernstein & Weiler, 2000), the cardinal feature of the child (Fletcher and Taylor, 1984) let alone all human life (Glozman, 2013) is development, there is little doubt that Vygotsky, with his intensive evaluation of the development of higher mental functions, is the founding parent of child neuropsychology.

Vygotsky was born in Belarus on November 17, 1896. He graduated from the University of Moscow in 1912 with an analysis of *Hamlet* as his dissertation (Ghassemzadeh, Posner, & Rothbart, 2013). He subsequently worked as a theater critic (van Der Veer, 2015) and as a teacher. Vygotsky's research in psychology started in 1924 (Ghassemzadeh et al., 2013). He was an insightful theorist and prolific writer. He died in 1934, after only 10 years in the field, with 270 manuscripts still left unpublished (Ghassemzadeh et al., 2013).

Any analysis of Vygotsky's theories is complicated not only by the sheer volume of his writings and the evolution in his thinking over time, but also by the likelihood that his work was suppressed by the Soviet government (Ghassemzadeh et al., 2013), censored, revised, intentionally falsified, and mistranslated (van Der Veer & Yasnitsky, 2011). Nonetheless, there is a consensus regarding the core elements of his theories. These refer directly to development, which we have argued elsewhere is an essential consideration in the interpretation of neuropsychological test data in children (Bernstein & Weiler, 2000).

Vygotsky is perhaps best known for his writings regarding the zone of proximal development, which continues to be referenced in current research in the areas of psychotherapy (Zonzi et al., 2014), coteaching (Murphy, Scantlebury, & Milne, 2015), gaming simulations (Clapper, 2015), and social-emotional and cognitive development in early childhood education (Bodrova, 2008). Although there are differences of opinion as to what Vygotsky actually meant (Murphy et al., 2015), we find the following explanations the most compelling: "functions which have not yet matured but are in the process of maturing . . . [are] . . . 'buds' or 'flowers' of development rather than 'fruits' of development" (Vygotsky, 1978, p. 86), and [the zone of proximal development represents] "the domain of transitions that are accessible by the child" (Vygotsky, 1987, p. 211).

Vygotsky was a contemporary of Piaget and Yerkes and used research to inform his theories. Prior to his research, many psychologists believed that mental operations were localized and essentially unchangeable. Vygotsky argued instead that these were much more complicated and variable systems that changed over the course of development and experience (Ghassemzadeh et al., 2013). “[Research] demonstrates, first, that no specific function is ever connected with the activity of one single brain center. It is always the product of the integral activity of strictly differentiated, hierarchically interconnected centers” (Vygotsky, 1997, p. 140). In other words, higher psychological functions are organized in a systemic fashion. They are not narrowly localized to one area of the brain. This provides the human brain with an amazing ability to adapt to the demands of the environment.

The second main tenant in Vygotsky’s theories is that functions are dynamically organized, and that the components involved in the completion of the higher psychological functions change over the course of development:

*In the process of development, and in the historical development of behavior in particular, it is not so much the functions that change as we studied before, (and it was our mistake), and it is also not so much their structure or the line of their development that changes. What is changed and modified are rather the relationships, the links between the functions. New constellations emerge which were unknown in the preceding stage. That is why intrafunctional change is often not essential in the transition from one stage to another. It is interfunctional changes, the changes of interfunctional connections and the interfunctional structure that matter. We will call the development of such new flexible relationships between functions a psychological system, giving it all the content that is usually attached to this, unfortunately, too broad concept.*

Vygotsky (1997), p. 92

The third idea, which is directly connected to Vygotsky’s cultural-historical approach to development, is that development occurs in the context of—and as a result of—social interactions. Social interactions are a necessary, but not sufficient, condition for development. Both nature and nurture have their roles to play. According to Vygotsky, higher psychological functions emerge first with the support of another individual before the child goes on to demonstrate the function independently.

*When we studied the processes of the higher functions in children we came to the following staggering conclusion: each higher form of behavior enters the scene twice in its development—first as a collective form of behavior, as an interpsychological function, then as an intrapsychological function, as a certain way of behaving. We do not notice this fact, because it is too commonplace and we are therefore blind to it.*

Vygotsky (1997), p. 95

With these core elements, Vygotsky places himself squarely inside the nature–nurture debate, arguing that both are necessary determinants of the child’s

behavior. Parenthetically, it is the lack of synchrony of these influences that is (1) genetically determined variability, and (2) social experiences occurring at different times for different children, that accounts for the diversity of atypical and typical developmental trajectories across children. Initially, biology drives development and limits the kinds of learning that take place. In subsequent stages, Vygotsky argued that learning drives development (Akhotina, 2003). Asynchronous biological development is the result of genetics. (On the nurture/social side, heterochrony is the result of the variety of interactions to which the child is exposed.)

Vygotsky differentiated between intrafunctional and interfunctional timetables of biological development. Intrafunctional development occurs when a specific system (e.g., the infant's visual system) matures, becoming more complicated, efficient, and automatic in its operation. Interfunctional development occurs as a result of changes in the components that interact in order to conduct the higher psychological function. Let us consider the emergence of the higher psychological function of *Curiosity* associated with understanding of and mastery over the environment. This is a drive that is seen in so many environments and with so many children, one might speculate that this is "hard-wired" into our neurology.

An infant begins to develop understanding of the world by first, responding to sensory (visual, auditory, tactile) stimuli and then orienting to the stimuli. When the visual system functionally matures at around 3–6 months (Bergen & Woodin, 2011), the grasping reflex, which has been present in an underutilized state since birth, has the conditions that will allow it to be deployed in a more purposeful way. The subsequent coordination and development of the visual and motor functions constitute the new visual–motor system (Glozman, 2013). This enables the child to express *Curiosity* about the environment in the form of a sensory evaluation of a self-chosen object. At the same time, other higher psychological functions (e.g., object permanence) are unfolding as a consequence of (1) their own genetically determined timetables and (2) the availability of conditions/experiences/new higher psychological functions that will enable them to emerge. This, we believe, is the essence of Vygotsky's zone of proximal development. Higher psychological functions can "bud," but only will bloom when the necessary environmental and social conditions are present.

For example, take that same infant who is tracking an object as it moves through space. An adult interacting with that child may remove the object from view, but bring it back to the child at a later time. At some point, the infant may have the insight that the disappearing and returning object is the same. If, at some point, the infant searches for an object that was hidden from view, we infer that he or she has developed object permanence (Bricker, Cap, & Pretti-Frontczak, 2002). With the development of object permanence, the child is now capable of understanding that objects (1) have characteristics, (2) behave in predictable ways, and (as the concept of causality emerges) (3) can be controlled.

The zone of proximal development includes aspects of both potential and timing. The parent's introduction of the word (or sign) for the child to use to label the object, can be useful to the child who is at the corresponding developmental stage. Adults name objects and label pictures in books throughout the child's early

development. With the emergence of the child's language system, the same object obtains a verbal label and the higher psychological function of *Curiosity* now incorporates the language function. As language develops (through social interactions) and, with additional experience, the child makes associations and places that object within a category, leading to a further understanding of the object.

Thus, over the course of development, new systems emerge to conduct the higher psychological functions and the distribution of these systems/networks change in a dynamic fashion. Because development of higher psychological functions is not limited by the maturation of any single brain component, there is both greater variability in and greater potential for development. In order to explain its development in the brain, it suffices to assume that the brain contains the conditions and possibilities for a combination of functions, a new synthesis, new systems, which do not at all have to be structurally engraved beforehand (Vygotsky, 1997, pp. 104–105).

From the perspective of a child neuropsychologist, Vygotsky's place as a founding parent of child neuropsychology comes from two factors: Vygotsky's insightful analysis of development interacted with his cultural-historical analysis to explain both the variety of typical patterns of development and the patterns of atypical development. The heterogeneity in the developmental timetables is genetically driven. These genetically derived timetables influence the differentiation, expansion, and increase in complexity within and between functions, and the timing of this can affect the way culture is experienced. As Glzman (2013) suggests, the maturation of the visual system changes the grasping reflex of the infant into a visual–motor operation. To expand this example further, this is then sensitive to the child's culture such as, for example, whether there are objects to grasp, and what variety there are.

## **Process of the child neuropsychological evaluation**

The purpose of the child neuropsychological evaluation drives the battery choice and design as well as the analytic approach. Some reviewers (e.g., Koziol & Budding, 2011) have argued persuasively for the use of strongly brain-based and brain-function based models and emerging research about the functioning of the brain to inform test selection. We agree with the importance of using information pertaining to the functioning of the brain to organize how one approaches an assessment. They cite, for example, the different functional pathways involved in retrieval from memory and recognition and therefore, the importance of enlisting tests of memory functioning that examine each of these separately. On the other hand, test interpretation and test-battery design are dictated by the end product. For example, in many instances, the purpose of the child neuropsychological assessment is (1) to provide a diagnosis that might enable a child to receive services at school, (2) to help identify interventions that would benefit the child, (3) to clarify future potential areas of risk, and (4) to enable the adults working with the child to better understand him or her.

Although one could certainly make the case that an anatomical-organization interpretation model could help identify future risk and boost understanding of the child, it is a yet unproven empirical question as to whether intervention recommendations derived from such an approach would be more effective than those derived from the research-validated interventions associated with specific disorders or ones that would be derived from a brain-function approach. For example, if the child struggled with access rather than retention of information, a potential modification might be multiple-choice testing rather than fill in the blanks and it is of little practical difference that one function may lie on an anterior–posterior axis whereas the other on a lateral–medial axis. Instead, our approach tries to be mindful of the interpretive and decision-making errors to which this kind of data analysis is vulnerable. To that end, we would advocate that an approach that disconfirms hypotheses is best suited to the ends we are desiring.

### ***The child neuropsychological versus psychoeducational evaluation***

A fundamental question to be addressed about purpose is related to what differentiates a child neuropsychological evaluation from the more common psychoeducational evaluation. Whereas individual assessments vary greatly depending upon the training and expertise of the examiner, there are specific differences between the two kinds of assessments. The first difference is the breadth of the evaluation. The main focus of a psychoeducational evaluation typically is to identify whether an individual has a learning disability or academic weakness that requires specific intervention. These evaluations are often carried out in schools and they provide estimates of an individual's aptitude as measured by standardized IQ tests, and one's academic achievement as measured by standardized academic tests. The scope of these evaluations is typically more limited as compared to the comprehensive evaluation of brain functioning offered by neuropsychological evaluations (Silver et al., 2006).

There are students for whom a standard psychoeducational evaluation may be sufficient; however, there are many other students for whom a standard evaluation for a learning disability is not sufficient. In contrast, neuropsychological evaluations tend to be broader in scope. Domains that are assessed are considered in light of the understanding of brain organization, brain–behavior relationships, and neurodevelopment. Neuropsychological evaluations cover a broad spectrum of issues and incorporate knowledge of brain functioning with observed performance.

Referral questions include evaluation for learning disabilities, attention disorders, language-processing disorders, nonverbal learning disabilities, Autism Spectrum disorders, and emotional disorders. The neuropsychological evaluation process includes obtaining extensive background information on the child's developmental history, medical history, school history, and socio-emotional functioning. As part of the assessment, information is obtained regarding not only the educational environment in which the child functions, but the social and family environment.

A detailed family history is taken in order to identify whether there is a family history of learning disabilities, psychiatric disorders, or other medical conditions. The evaluation is tailored to the referral question and suspected concerns. Neuropsychological evaluations can include standardized IQ tests, measures of memory, language processing, visual–spatial abilities, executive functions and attentional processes, academic achievement, visual–motor skills, mental-processing speed, and social-emotional ratings. Children who are struggling academically may do so for a variety of reasons, and a neuropsychological evaluation can help parents and educators understand what is happening, why it is occurring, and how to intervene.

The second difference between the two kinds of evaluation is the examiner's training and experience. Psychoeducational assessments are often conducted by professionals in the schools. The school psychologist will usually conduct an evaluation including aptitude testing. The educational specialist will complete standardized educational testing. These evaluations are discussed and compared to target what services a child may require. Neuropsychologists are psychologists who have specialized training in understanding brain–behavior relationships. They have completed appropriate course work in neuropsychology, internship training, and a post-doctoral fellowship. Their training allows them to apply an understanding of psychology, physiology, and neurology to complete a comprehensive assessment, which includes diagnoses and recommendations.

Educational recommendations from psychoeducational assessments are developed that identify strategies to intervene or compensate for areas of academic weakness and ways to capitalize on the individual's relative academic strengths. Recommendations from these evaluations are typically classroom-based. When a child has a neuropsychological evaluation, recommendations are developed based on an understanding of academic strengths and weaknesses as well as a neurocognitive processing profile and developmental stage. This allows one to develop appropriate academic, social-emotional, and behavioral recommendations.

Perhaps the biggest difference separating the neuropsychological assessment of the child population from other types of cognitive or psychoeducational assessments is the degree to which the neuropsychological evaluation must take into account not only the various cognitive operations and processing systems involved in completing the tasks but also the way in which the child's developmental course and social and medical history has brought him or her to this particular point in time. It is through the consideration of these multiple factors that the child neuropsychologist attempts to try to explain not only why the child is unable to do particular kinds of everyday tasks that come so easily to other children of the same age, but also why the child is able to do other kinds of tasks.

In some cases, providing an explanation for the child's behavior is helpful in and of itself (e.g., the child is not doing what one is asking because the child does not understand the language one is using, not that the child is being disrespectful or disobedient). It can provide those adults working with the child an alternative narrative that assists them in being more patient and effective. More importantly, the understanding derived from the comprehensive assessment enables the clinician to make informed recommendations about the interventions that will work the best with the

child and reasonable predictions about what obstacles the child is likely to face at future developmental stages. In many cases, this is achieved by an understanding of the disease, disability, or injury affecting the child and making predictions based on the course and later risk factors frequently observed in similar children. In other cases, when the child's abilities and disabilities align less clearly with any one disorder domain, the predictions have a less sound foundation and are created based on an understanding of typical development, how the child dealt with prior developmental obstacles, and how he or she is likely to respond to subsequent ones. At other times, when the issues confronting the child are less clear, it can be the accuracy of these predictions (i.e., how the child manages the subsequent stages of academic, social, or emotional development) that supports or disconfirms the diagnosis.

This comprehensive, developmental view of the child is also what can differentiate the child neuropsychological assessment from the neuropsychological assessment of adults in that the latter assessment is most often requested to understand the loss of function in a fully developed adult whereas the former is typically sought to understand the individual who is off developmental track and has not obtained the milestones expected of him or her. This, in essence, is what the training of child neuropsychologists is intended to comprise: normal development of the child, the ways in which historical factors (e.g., prenatal history, early exposure, injury, illness) affect development, the ways in which environmental factors (e.g., poverty, lead exposure, adequate schooling) affect development, what tests measure and what contributes to error, how observations and questionnaires can add to or reduce understanding of the child, and how to use an analytical/diagnostic system to make sense of the large amount of information derived during the assessment.

### ***Theory and goals of the assessment process—deriving meaning***

The meaning of the neuropsychological evaluation is derived from its ability to answer the question that prompted the evaluation. In the past, a common question posed to neuropsychologists and asked of their assessment battery was, "Are there signs of brain injury ([Nici & Reitan, 1986](#)) and, if so, where is its location ([Kertesz, 1984](#); [Wilkening, 1989](#))?" With the advent of newer and more reliable tools for identifying and localizing structural and metabolic abnormalities ([Olson & Jacobson, 2015](#)), and with an increase in the range of children referred for evaluation, some of the referral questions have changed. In contemporary child neuropsychology practice, we are asked a range of referral questions, but the majority of these questions emerge in the form of, "What is causing this child not to be at ease in the environment?" This is the question that is often related but not necessarily identical to the initial referral question.

To answer this question and to be able to generate recommendations that will reduce the child's suffering and improve his or her adaptation to the environment it is necessary to include interpretable test results within an understanding that the child's current behavioral repertoire is the outcome of the child's developmental course to date ([Bernstein & Weiler, 2000](#)). These fall within the larger tableau of the critical contributing factors: (1) culture (in the Vygotskian sense of, language,

environment, resources/practices, presence of poverty, health conditions, family resources, community, schooling, potential opportunities); (2) experience (historical) factors including parenting practices/parental availability, opportunities for schooling, past experiences with schooling, social opportunities; and (3) development in the Vygotskian sense of the interplay between genetics and timing. In the zone of proximal development, the bud may be ready but will only bloom if the necessary conditions are provided.

The additional challenge is that as the child ages, the demands of the contexts within which the child operates (e.g., family: inter-relationships and the child's desire for increased autonomy; school: academic, organization and independence expectations; peer: language and social functioning) will continue to change ([Bernstein & Weiler, 2000](#)). Genetically determined brain changes will occur with new developmental zones budding (and given the correct environment, new psychological functions will bloom). As a result, the child's future path, even at the end of the most thorough assessment, will still be far from certain.

The majority of children referred for evaluation are at increased risk. Brain dysfunction decreases a child's ability to adapt, which will directly or indirectly increase the risk for emotional and behavioral problems. These occur as a result of increased exposure to failure and frustration, changes in temperament and personality development, family responses, the child's reaction to the disability, and the possible negative effects from the treatments and interventions themselves ([Tramontana & Hooper, 1997](#)). An accurate assessment can start the process whereby the risk of subsequent development of psychiatric problems is reduced ([Tramontana, 1983](#)). Nonetheless, even life-saving interventions can put the child off his or her typical developmental course (see, e.g., Bernstein's exposition of myelodysplasia in [Bernstein & Weiler, 2000](#)).

There are multiple avenues through which the neuropsychological evaluation can be beneficial to the child. First, the evaluation can be structured in a way that enables the child to feel successful. For example, by attending to rapport and the quality of the adult-child interaction, managing the feedback to the child, and modifying the order, expectations, and interpretations and attributions the child makes to challenging tasks, it is likely that the child will come away from the evaluation having had a positive interaction with a supportive adult. Similarly, explaining and redefining the functions of the concerning behaviors to the parents can increase their tolerance, reduce scapegoating ([Tramontana & Hooper, 1997](#)), and improve their ability to advocate for their child in the future.

School serves as the child's most important social and learning environment outside of the home and social and academic success improves the quality of a child's life ([Fletcher-Janzen, 2005](#)). Therefore, increasing the child's success at school is a meaningful goal for the evaluation. One of the most important targets for the neuropsychological evaluation is the narrative the school personnel use to explain the child's behaviors. Adults at home and at school may characterize the child as being unmotivated, disagreeable, oppositional, or stupid. If they view the child as lazy, apathetic, or otherwise difficult, they will generate expectations that will perpetuate the child's existing problems. Both positive and negative narratives can create expectations and self-fulfilling prophecies.

Recommendations and interventions, which are inherent to the neuropsychological evaluation, may point toward strategies that can remediate the deficits. Recommendations are probabilistic endeavors because there is error inherent in the tests, uncertainty how the recommendations will be delivered, and unknowns regarding how the child will react. For these reasons, an understanding of evidence-based intervention strategies is critical as at least a starting point. In this regard, the proposed interventions must be valid within the child's environment. This reinforces the importance of having ecologically valid assessments that support the capacity of the evaluation to confirm, validate, and interpret the referral concerns. When there has been limited response to these interventions, an evaluation of the child's profile as well as an understanding of brain–behavior relationships may allow the neuropsychologist to generate alternative approaches.

Complicating the diagnostic and intervention process is that as children fall further behind their classmates, it becomes more difficult for them to catch up and each additional failure erodes self-confidence and motivation (Tramontana & Hooper, 1997). The neuropsychological evaluation can serve as a “gatekeeper” operation through which the child is able to obtain access to additional services and supports. It is in this context as well as the research based recommendations that result, that diagnosis is so important.

*Assessment approaches.* A neuropsychological evaluation is a process rather than a product (Fletcher-Janzen, 2005) and it is an evolving process as it regularly is called upon to adapt to new knowledge about the child and development, new diagnostic tools, pathophysiology, and the developmental progression of disorders and diseases. The focus of the larger question is not so much on what is wrong but on how the child meets the demands (both successfully and unsuccessfully) of his or her environment (Bernstein & Weiler, 2000).

There are a variety of authors who explore the neuropsychological evaluation in great detail (e.g., Fletcher-Janzen, 2005) and who discuss the strengths and weakness of fixed, flexible, and transactional test batteries. Our approach, in which we try to disconfirm hypotheses for the child's poor adaptation, tends to favor a more fixed type of battery. The majority of test batteries reviewed by Fletcher-Janzen (2005) include within their assessments tests of intelligence, memory, learning, language, motor skills, visual–spatial functions, frontal executive (including attention), academic achievement, social-emotional functioning (personality, behavior) and a few additional domains such as body awareness, sequential processing, psychosocial factors, environment fit, sensory–perceptual skills, problem-solving, and concept formation. Again, for us, the appropriate battery is one that allows one to rule out possible explanations for why the child is struggling with the idea that one or more of the remaining hypotheses should be true.

*Research design and methodology.* The clinical assessment is a procedure for answering specific questions about behavior, in this case, that of a given individual. It is formally equivalent to the traditional research study (see also Pennington, 1991; Rabinowitz, 1994). Thus, it is our position that the clinical assessment must be treated as an experiment with an  $n$  of 1. Both the experiment with the  $n$  of  $N$  (the research experiment) and the experiment with the  $n$  of 1 (the clinical

experiment) seek to answer research (diagnostic) questions by disconfirming hypotheses involving observed phenomena. The answers to research questions are not derived directly from the statistical analysis of the data; interpretation is necessary to understand the meaning of the data (Kerlinger, 1986). Similarly, test scores alone are not sufficient for diagnosis (let alone intervention). The research or diagnostic question is answered by the design of the research study (assessment). A carefully designed assessment enables conclusions to be inferred; without such a design, alternative hypotheses cannot be excluded. We believe that the extensive knowledge base available in research methodology and design can (and should) be applied explicitly to the challenge of assessment. It is through the application of  $n$  of  $N$  procedures to the  $n$  of 1 environment that best practices for child neuropsychological assessment (and assessment in general) can be derived.

As Kerlinger (1986) states, three criteria for judging the effectiveness of research designs are: (1) adequate testing of hypotheses, (2) control of variance, and (3) generalizability. For the neuropsychological evaluation, adequate testing of hypotheses refers to the accumulation of information that will allow relevant hypotheses to be disconfirmed. Threats to construct validity, which operate in the research study, are equally of concern in the  $n$  of 1 clinical experiment. In order for a design to be able to test a hypothesis truly, data must be collected in such a way that it can be falsified if it is, in fact, incorrect (Rosenthal & Rosnow, 1984). In the experimental design, variance control pertains to three separate types of variance. Systemic variance related to the experimental manipulation should be maximized. The effects of variance related to factors extraneous to the experimental manipulation should be controlled through the use of random assignment or incorporation of the factors in the research analysis. Error variance is minimized through standardized procedures and selection of reliable measurement tools.

In the neuropsychological evaluation, maximization of systemic variance can be achieved by using tools and techniques that are sufficiently sensitive to identify individuals unambiguously with the diagnoses of concern. Control of variance related to factors extraneous to the diagnosis can be achieved through inclusion of factors (e.g., developmental history, academic exposure, emotional trauma) in the overall understanding of the child (i.e., the child in relation to his or her systems) and reduction of other variables (e.g., hunger, fatigue, boredom, fear, etc.) that can invalidate the testing results. As in research designs, error variance is minimized through the use and standardized application of reliable measurement tools and consistent (replicable) data integration procedures.

In research designs, the generalizability standard (one of the various forms of validity) requires that the results of the study be applicable to other groups in other situations. In the neuropsychological assessment, generalizability refers to the likelihood of the child's performance during the evaluation being relevant to the child's behavior in his or her real-life environment (ecological validity).

Research design constitutes a multiply redundant system in which formal procedures are applied at different levels. These include: the original theory (theories); the generation of hypotheses; the experimental design; group selection; task design; task administration (formal procedures); and reliability in administration, in scoring,

and in the collection of normative data. The clinical “experiment” entails different responses to the requirements of research design and method than that of the research investigation. For example, the clinical experiment cannot conform to the assumptions of parametric statistics, nor can it eliminate the preconceptions and already acquired knowledge of the clinician—making a Bayesian analysis necessary (Murphy, 1979). Where formal strategies are not available, as in the case of Bayesian methodology, however, the same goal of experimental control may require alternative applications of methodological tools/techniques. For example, in the experimental setting the influence of the observer can be subjected to the same control as any other variable, that is, it is “averaged” across multiple observations, of the same behavior by multiple persons. The impact of more than one observer can, if needed, be formally tested post-hoc by directly comparing the distributions of observations made by one observer with those of another. In the clinical—experimental setting, this is not possible: clinical assessment is done with one person at a time. Alternative methodological strategies thus must be employed to obtain the same goal of experimental control. Multiple observations from multiple informants (i.e., a multimethod design; Campbell & Fiske, 1959) are still necessary. These are, however, necessarily different. To be of value in diagnosis, they must be congruent in their implications: They must both converge on, and discriminate among, coherent diagnostic entities. It is the theory of the child, not the theory of assessment or of the tools, that determines the congruence; the critical validities for diagnosis are convergent and discriminant (see also Pennington, 1991). Methodologically, two conditions must be met: (1) converging data must be derived from multiple domains (a cross-domain analysis); and (2) behaviors that are not predicted by the theory for the neural substrate in question should not be present. Note that the latter must be sought actively and systematically to counter the risk of confirmatory bias.

For a majority of neuropsychologists, an IQ test is an essential part of the neuropsychological evaluation because it is one of the essential elements in the neuropsychologist’s role as a gatekeeper when working with schools and testing agencies such as the College Board. Therefore, it is important to understand the thinking behind the design of current IQ tests.

**CHC theory.** The Cattell–Horn–Carroll (CHC) theory of cognitive abilities (see Schneider & McGrew, 2012) shows great potential for providing a framework for the child neuropsychological evaluation of cognitive functions. The theory represents an integration of two major, well supported conceptualizations, that is, Carroll’s three-stratum theory (Carroll, 2012) and extensions of the Horn–Cattell theory of fluid (Gf) and crystallized (Gc) intelligence (Horn & Blankson, 2012). As Schneider and McGrew attest, a great deal of consensus has accrued over the past 25 years converging on CHC theory as a typology of human cognitive abilities, and we believe that contemporary child neuropsychologists are well advised to consider this typology in conceptualizing the neuropsychological evaluation and in choosing assessment instruments (see also, Brinkman, Decker, & Dean, 2005). At the same time, we must recognize the *potential* value of the theory, because although the structural (i.e., factor analytic) and developmental basis of the theory has been well established, its neurological basis and neuropsychological treatment validity and

utility have yet to be studied nearly as extensively. Indeed, when tests based on this theory have been assessed for their incremental validity in predicting academic achievement, lower order composite factors designed to measure particular CHC constructs have not fared especially well in comparison to their higher order full-scale counterparts (see, e.g., McGill & Busse, 2015; McGill, 2015). It is clear, however, that the assessment of neurocognitive functions is central to the child neuropsychological evaluation and that most contemporary individually administered tests of intelligence are either based on or at least reference CHC theory (Brinkman et al., 2005; Keith & Reynolds, 2010; Willis, 2017). Indeed, based on their review of research, Keith and Reynolds noted that factor analyses of contemporary individually administered intelligence tests supports CHC constructs as the latent traits underlying those tests.

According to Horn and Blankson (2012), about 80 first-order human-cognitive or primary-mental abilities have been identified through factor-analytic research (Carroll, 1993; Horn, 1991), and the intercorrelations among those factors can be reduced to about eight broader components or second-order factors. Moreover, those second-order abilities also are correlated to form a hierarchical organization of human cognitive functions. In contrast to an overriding Spearman's *g*, however, Horn and Blankson emphasize that developmental and neurological evidence militates against a single factor, but instead suggests three, third-order clusters of abilities: that is (1) vulnerable abilities, (2) expertise abilities, and (3) sensory–perceptual abilities. Vulnerable abilities are characterized by their declines in function associated with age in adulthood and neurological deterioration. In contrast, expertise abilities are characterized by their improvement, maintenance, and resistance to aging declines in adulthood. Sensory–perceptual abilities do not fall clearly into either of the other two clusters; instead, they do not typically diminish as early, as regularly, or as much as vulnerable abilities, and yet do not improve as consistently or as much as expertise abilities (Horn & Blankson, 2012). Carroll (2012) postulated fewer than ten second-stratum or second-order abilities (comprising about 65 narrow or first-stratum abilities; see Carroll, 1993), all of which are subsumed by a highest-order, or third-stratum general factor, that is, *g*. Table 7.1 is adapted from Horn and Blankson to show the relationships among eight second-order abilities and their corresponding third-order conceptual clusters. Table 7.2 is adapted from Carroll (2012) listing his broad, stratum II cognitive abilities.

Schneider and McGrew (2012) have elaborated on these conceptualizations and have postulated no less than 16 second-order factors as forming the basis of human cognitive functioning in their hybrid CHC theory. They also organize the second-order factors according to three third-order clusters that closely resemble the developmental organization proposed by Horn and Blankson (2012). Table 7.3 is adapted from Schneider and McGrew to show the relationships among 16 CHC second-order abilities and their corresponding third-order clusters. As noted, Keith and Reynolds (2010) reviewed confirmatory factor-analytic data on 13 new and revised tests of intelligence (i.e., Woodcock–Johnson Psychoeducational Battery-Revised, Woodcock–Johnson III Tests of Cognitive Abilities, Differential Ability Scales,

**Table 7.1** Relationships among eight Horn–Cattell second-order factors and third-order clusters

Third-order cluster	Label	Description
Vulnerable abilities	Gf	(Fluid) abilities of reasoning under novel conditions
	SAR	Abilities of short-term apprehension and retrieval
	Gs	Speed of thinking abilities
	Gc	(Crystallized) acculturation knowledge abilities
Expertise abilities	TSR	Abilities of long-term (tertiary) storage and retrieval
	Gq	Quantitative and mathematical abilities
Sensory–perceptual abilities	Gv	Visualization and spatial orientation abilities
	Ga	(Auditory) abilities of listening and hearing

Source: Adapted from Horn, J. L., & Blankson, N. (2012). Foundations for a better understanding of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed., pp. 73–98). New York: Guilford Press.

**Table 7.2** List of Carroll's broad (stratum II) cognitive abilities

Fluid intelligence
Crystallized intelligence
General learning and memory
Broad visual perception
Broad auditory perception
Broad retrieval ability
Broad cognitive speediness
Processing speed (reaction time decision speed)

Source: Adapted from Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press.

Detroit Tests of Learning Aptitude-3, Kaufman Assessment Battery for Children, Kaufman Assessment Battery for Children-Second Edition, Kaufman Adolescent and Adult Intelligence Test, Stanford–Binet Intelligence Scales—fourth edition, Stanford–Binet Intelligence Scales—fifth edition, Wechsler Intelligence Scale for Children—third edition, Wechsler Intelligence Scale for Children—fourth edition, Differential Ability Scales, and Cognitive Assessment System), and found general support for the CHC model across a wide span of instruments. Notably, most of these analyses showed evidence of substantially fewer second-order abilities than postulated by Schneider and McGrew.

Carroll (2012) emphasized that his theory provides a “map” (p. 887) of currently known human cognitive abilities, but warns that scores on most presently used assessment instruments reflect multiple factors at different strata. Thus, although attempts have been made to match particular second-order factors or broad stratum-II abilities to particular cognitive and neuropsychological tests and subtests (see Flanagan, Alfonso, & Ortiz, 2012; Miller & Maricle, 2012), such measures are

**Table 7.3** Relationships among 16 CHC second-order factors and third-order clusters

<b>Cluster</b>	<b>Label</b>	<b>Description</b>	<b>Definition</b>
Domain-free general capacities	Gf	Fluid reasoning	Solve unfamiliar problems
	Gsm	Short-term memory	Encode, maintain, and manipulate information in immediate awareness
	Glr	Long-term storage and retrieval	Store, consolidate, and retrieve information over periods of time
	Gs	Processing speed	Perform simple, repetitive cognitive tasks quickly and fluently
	Gt	Reaction and decision speed	Speed of simple judgments when presented one at a time
	Gps	Psychomotor speed	Speed and fluidity of physical body movements
	Gc	Comprehension-knowledge	Depth and breadth of knowledge and skills valued by culture
	Gkn	Domain-specific knowledge	Depth, breadth, and mastery of specialized knowledge
	Grw	Reading and writing	Depth and breadth of knowledge and skills related to written language
	Gq	Quantitative knowledge	Depth and breadth of knowledge related to mathematics
Acquired knowledge	Gv	Visual processing	Use of simulated mental imagery to solve problems
	Ga	Auditory processing	Detect and process meaningful nonverbal information in sound
	Go	Olfactory abilities	Detect and process meaningful information in odors
	Gh	Tactile (haptic) abilities	Detect and process meaningful nonverbal information in touch sensations
	Gk	Kinesthetic abilities	Detect and process meaningful nonverbal information in proprioceptive sensations
	Gp	Psychomotor abilities	Perform physical body movements with precision, coordination, or strength
Sensory- and motor-linked abilities	Gv	Visual processing	Use of simulated mental imagery to solve problems
	Ga	Auditory processing	Detect and process meaningful nonverbal information in sound
	Go	Olfactory abilities	Detect and process meaningful information in odors
	Gh	Tactile (haptic) abilities	Detect and process meaningful nonverbal information in touch sensations
	Gk	Kinesthetic abilities	Detect and process meaningful nonverbal information in proprioceptive sensations
	Gp	Psychomotor abilities	Perform physical body movements with precision, coordination, or strength

Source: Adapted from Schneider, W. J., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed., pp. 99–144). New York: Guilford Press.

unlikely to yield true assessments of latent neuropsychological constructs postulated in CHC theory. Moreover, as Carroll has discussed, assessing a child on the total range of abilities postulated by these theories is unrealistic given time constraints even if well validated measures were available for all of these human cognitive

abilities. Perhaps, as [Miller and Maricle \(2012\)](#) suggest, contemporary intelligence tests, which are largely based on CHC theory, might be viewed as an initial step in generating hypotheses about potential neurocognitive constructs to be followed up with more specialized neuropsychological measures. At the same time, these kinds of cross-battery assessment methods must be approached cautiously given psychometric concerns about profile analysis, standard errors of difference scores, typical base rates of intersubtest scatter, and ecological and incremental validity.

## Sources of error

### *Incremental validity*

[Hunsley \(2003, p. 443\)](#) described incremental validity as the “extent to which a measure adds to the prediction of a criterion beyond what can be predicted with other data.” This often-neglected form of validity refers to the degree of accuracy that a particular source of information can add to a decision above and beyond the degree of accuracy already achieved by other measures. As [McGill and Busse \(2015\)](#) note, incremental validity has, at its core, the law of parsimony, and when applied to neurocognitive batteries for children, this implies that interpretation ought to occur at the broadest and most reliable ability level, unless there is a compelling reason to do otherwise. Incremental validity varies not only as a function of the source of information added, but also as a function of the kind of decision being considered. Differential neuropsychological decisions relate to differential goals including diagnosis, treatment planning, treatment evaluation, and progress monitoring ([Johnston & Murray, 2003](#)), and therefore incremental validity differs as a function of these kinds of decisions. For example, adding a teacher’s report of behavior to a battery of neuropsychological tests for a child may show limited incremental validity for the treatment of Attention-deficit Hyperactivity Disorder (ADHD), but tremendous incremental validity for the diagnosis of that disorder. Similarly, a behavioral observation of a child in a classroom may improve the accuracy of a diagnosis made in a physician’s office, but add nothing unique to a teacher’s behavioral rating. Many child neuropsychological tests show relatively poor levels of incremental validity when compared to the diagnostic power of contemporary neuroimaging techniques and yet still are important because they can contribute useful information about current levels of neurocognitive functioning.

It is important, however, for neuropsychologists to avoid over estimating the degree to which particular sources of data may improve upon the validity of information that already has been collected. For example, a common practice among many psychologists is to conduct an analysis of a particular profile of subtests based on a neurocognitive battery of tests. This kind of profile analysis of subtest scores, however, is fraught with psychometric concerns and even broader indexes or factor scores that have been supported through confirmatory factor-analytic statistical approaches recently have been shown to lack incremental treatment validity

beyond full-scale composites (Glutting, Watkins, Konold, & McDermott, 2006; McGill, 2015; McGill & Busse, 2015).

First of all, one must consider the standard error of the difference between scores when attempting to understand the way in which a set of subtest scores may fluctuate for an individual child. This requires one to consider the reliabilities of both scores that contribute to the difference. Although it is tempting to assume that the reliability of a difference score exceeds that of either of the scores that contribute to it owing to the number of items involved, the opposite actually is true. Here, the reliability of a difference score is lower than the reliability of either of the scores contributing to it because a standard error of a difference (SED) equals the square root of the sum of the squares of the two standard errors of measurement (SEM; which are a function of both reliability and standard deviation) for the two subtests, that is,  $SED = \sqrt{SEM_1^2 + SEM_2^2}$ . For example, assume that the standard errors of measurement for the subtests are 3 and 4 points:  $SED = \sqrt{3^2 + 4^2} = 5$  points, which, of course, exceeds either of the standard errors of measurement of the subtest scores compared.

Even when subtest-score differences are reliable, meaning that they exceed the amount of difference expected by chance, and that the difference is likely to be replicated upon retest, we still need to examine the prevalence, or base rate, of that difference within the population of interest. Fortunately, this is relatively easy to do, given most contemporary examiner and technical manuals that accompany major cognitive tests contain this information. Here, one must examine the cumulative frequencies in the normative sample between various combinations of subtests to determine how unusual they may be within the normal population. Test users often are surprised to discover that even very high degrees of intersubtest scatter, that is, the difference between the highest and lowest subtest scaled scores, occur quite commonly within the standardization samples for most of the major neurocognitive batteries available today. Perhaps this occurs because, as psychologists with our statistical training, we refer to these reliable differences as *significant*, and it is easy to confuse significance with clinical meaningfulness. For example, the average amount of intersubtest scatter on the Wechsler Intelligence Scale for Children, fifth edition (WISC-V: Wechsler, 2014) is 7 scaled-score points, and a three standard-deviation difference (i.e., 9 scaled-score points) occurs in 23% of the standardization sample for the 10 primary subtests (Wechsler, 2014, p. 367). Of course, what constitutes an unusual difference is a subjective criterion, although few would argue that nearly one quarter of the population represents an unusual degree of difference.

Finally, even in the situation where subtest-score differences are both reliable and unusual, one must consider the issue of incremental validity. Does knowledge of a replicable difference with a low base-rate of occurrence improve our understanding of the neurocognitive pattern of strengths and weaknesses? Again, we must access our understanding of the incremental value of this level of analysis. As McGill (2015) summarizes, the studies of incremental validity using tests based on CHC theory have used academic achievement as the criterion against which to compare broad secondary factors. Practitioners who focus on specific neurocognitive

domains, however, may be more interested in alternative criteria and, as noted previously, incremental validity varies as a function of the kind of decision being considered, such as, for example, the diagnosis of a particular neurocognitive disorder (e.g., Decker, Hale, & Flanagan, 2013; Keith, 1994).

### ***Culture, language, and poverty***

Still other sources of error include failure to properly account for variables such as the effects of ethnicity, race, language, and poverty. The extent of ethnic diversity among the children who are referred for evaluation is likely to be considerably influenced by location of the practice. For example, in a heterogeneous metropolitan area such as New York City, where over 200 languages are spoken, the home language for a third of the population is not English (United States Census, 2015). Across the entire country 21% of the population speak a language other than English in the home (United States Census, 2013).

There is disagreement as to how this population heterogeneity should even be summarized. Although the terms race and ethnicity are sometimes used synonymously, race infers to greater genetic similarity whereas ethnicity refers to characteristics such as language and national origin (Olson & Jacobson, 2015). Recent research suggests that race does not, in fact, reliably identify true genetic differences (Brickman, Cabo, & Manly, 2006; Olson & Jacobson, 2015) nor should we assume that all individuals who self-identify as, for example, Hispanic have the same cultural experience regardless of where their families originated from or where in the United States they might live (Olson & Jacobson, 2015).

Error can be introduced to the assessment if tests used to assess children from different backgrounds (e.g., ethnicity, race, culture, or socioeconomic status) are not measuring exactly the same construct in all children. For example, even after controlling for other variables, reliable differences among ethnic groups have been found (Boone, Victor, Wen, Razani, & Ponton, 2007; Brickman et al., 2006). Clearly, understanding the important role that experience plays in development, Vygotsky would not have been surprised by this result. Nonetheless, among other concerns, this finding raises the question as to what is actually captured by these categories. Some have argued that the categories of race and ethnicity are markers of other factors that could affect brain functioning such as socioeconomic status, quality of education, and acculturation (Olson & Jacobson, 2015). Socioeconomic status has been shown to be associated with test performance particularly among young children (Olson & Jacobson, 2015) through intermediaries such as education, home environment, availability of books, time available to read to child, parental education, employment, nutrition, and access to healthcare (Brickman et al., 2006). Cultural differences might affect the living environment, the child-rearing philosophy, and the attitude of the child's caregiver toward child development (Olson & Jacobson, 2015). It also would be incorrect to assume that test-taking skills such as working quickly are not affected by the child's culture or relationship between the child and the examiner (Olson & Jacobson, 2015).

Although nonverbal tests, because of their reduced language demands were once thought to be “safer” for children from different cultural backgrounds, we have since learned that education and cultural background influence a child’s exposure to drawing tasks and that parental education and child-rearing practices can influence a child’s level of attention and executive skills (Rosselli & Ardila, 2003). Similarly, it would be incorrect to assume that test results would be similar if the test were administered in the child’s native language (Brickman, et al., 2006). Translation does not ensure that the translated word is equally familiar or equally easy to remember in the child’s native language (Olson & Jacobson, 2015).

Tests developed in the United States attempt to create their standardization sample to closely match that of the U.S. population by race, ethnicity, parental education, and geographic region. But the question remains as to whether this, in fact, solves the problem. Manly (2008, p.179) argued the opposite, “There is widespread agreement that many neuropsychological measures do not have acceptable diagnostic accuracy when used among people who are not Caucasian, well-educated, native English-speaking, and middle to upper class.” The use of race-specific norms would not necessarily be beneficial as they would not account for either the diversity within the race category or the construct validity of the test itself and it would be logically impossible to create norms for the wide variety of ethnic groups (Brickman et al., 2006).

Because of their exposure to the culturally dominant language in school, some of the children seen for child neuropsychological evaluations will be bilingual. Their appropriate normative group would not only have to account for native language and country of origin, but also factors such as length of time in each culture and amount of (and level of proficiency of) exposure to two languages. The latter is of critical importance because it is thought that second-language learning progresses through multiple stages, some of which include a period of silence and a period of language loss (Jontak, 2013). Bilingual exposure appears to confer both advantages and challenges to the child. There is research that suggests that bilinguals have an advantage over monolinguals on tests of selective attention (Genesee, 2015). At the same time, although their conceptual vocabulary (i.e., both languages considered together) is similar when each language is considered separately, their vocabularies are smaller and they may not achieve a similar level of language competence as their monolingual peers (Genesee, 2015).

Clearly, child neuropsychologists must be culturally competent. Mindt, Byrd, Saez, and Manly (2010) noted that this entails (1) continuous awareness by the clinician of his or her assumptions about human behavior, values, biases and stereotypes, and how these might negatively affect the neuropsychological evaluation process as well as the test outcomes; (2) active efforts by the clinician to understand the worldview of culturally dissimilar clients without criticizing or passing negative judgments; and (3) ongoing and intentional efforts by the clinician to develop and practice relevant and sensitive assessment techniques and communication skills when working with ethnically diverse clients.

Our view is that neuropsychological evaluation of children from culturally diverse communities is a reality because some of these children will be in need of

supportive intervention. Evaluations will be necessary despite the inadequacy of the norms, questions about the constructs the tests are measuring, and uncertainty as to how the child's prior experiences have affected his or her performance. Neuropsychologists with excellent understandings of the child's culture will reduce, but not completely eliminate, all of these sources of error, leaving them in a position where a degree of caution is required in making recommendations and predictions. In the end, when confronted with a lack of research-validated decision-making rules we agree with Meehl (1957, p. 269) who suggested that "we use our heads because there isn't anything else to use."

### ***Ecological validity***

Although somewhat changed from its initial use, ecological validity is now commonly used to indicate the apparent similarity of the test and the real-world construct it is attempting to measure and the degree to which performance on the test predicts real-world behaviors (Burgess et al., 2006). Verisimilitude represents the extent to which the cognitive demands of the test appear to reflect the cognitive demands of the environment, whereas veridicality describes the demonstrated empirical relationship between the test and other functional measures (Franzen & Wilhelm, 1996).

Tests vary in their ecological validity, and some of this may be related to their origin. Tests derived from experimental paradigms were likely developed to answer a specific question about brain–behavior relationships and were dependent upon the extant knowledge about what the test was to measure and how that was to be accomplished. Very often these original conceptions have been modified but the original test remains. For example, the Stroop was developed in 1935, the Wisconsin Card Sorting Test in 1948, and the Tower of London in 1982 (Burgess et al., 2006). It is important to keep in mind that even small changes in test format can cause large changes in task demands (Stuss, Binns, Murphy, & Alexander, 2002).

The testing process itself also limits ecological validity because, by nature, we sample a small selection of the child's behavior in a single environment, without the emotional and behavioral challenges of the child's everyday life (Chaytor & Schmitter-Edgecombe, 2003). A few studies have broadly demonstrated an empirical relationship between tests of attention and real-world measures of adaptive functioning (Price, Joscko, & Kerns, 2003). On the other hand, a 2003 review (Chaytor & Schmitter-Edgecombe, 2003) of neuropsychological test results and real-world skills suggested only moderate to weak relationships between the two. An alternative approach, that has yielded tests with good psychometric properties, has been to design a neuropsychological test directly from the individual's real-world demands, for example, planning a route to complete multiple errands on a shopping trip (Knight, Alderman, & Burgess, 2002). In any event, it should never be assumed how well one can generalize from a single observation to the entire domain of potential observations if the examiner had unlimited time (Cronbach, Rajarathnam, & Gleser, 1963).

## ***Malingering***

The validity of any child neuropsychological evaluation is contingent upon getting a valid report of an examinee's symptoms and the examinee exhibiting optimal effort during testing. Larrabee (2012) suggested the term symptom validity testing when referring to the assessment of whether someone is exaggerating or fabricating self-reported symptoms, and the term performance validity testing referring to the use of specific measures to detect suboptimal effort or malingering on cognitive tests. In order to assess effort, neuropsychological evaluations have increasingly included performance validity testing in their battery.

Although there are many reasons why an individual may not exert optimal effort, malingering refers to the intentional exaggeration of impairments. Malingering is the intentional production of false or grossly exaggerated physical or psychological symptoms motivated by external incentives. Prior to the 1980s very little performance validity research was conducted; however, with the proliferation of neuropsychology into the forensic area, results and methods were scrutinized more closely. In 2007, the American Academy for Clinical Neuropsychology issued a Practice Guideline for Neuropsychological Assessment and Consultation (Heilbronner et al., 2009): "...the assessment of effort and motivation is important in any clinical setting, as a patient's effort may be compromised even in the absence of any potential or active litigation, compensation, or financial incentives. Clinicians utilize multiple indicators of effort, including tasks and paradigms validated for this purpose, to ensure that decisions regarding adequacy of effort are based on converging evidence from several sources, rather than depending on a single measure or method" (p. 221–222).

Although there has been a great deal of research over the past two decades focusing on malingering and suboptimal effort in adults, research in pediatric performance validity testing has been lacking, in part, because children were considered less likely and less able to deceive than adults (Salekin, Kubak, & Lee, 2008). Motivation to feign impairments was also less clear in a pediatric population. Just as in an adult population, however, there are many reasons why a child may exhibit suboptimal effort, including psychological factors, school avoidance, modifications/time extensions on standardized testing, and social factors (Kirkwood, Kirk, Blaha, & Wilson, 2010).

Although at one point it was a common belief that children would not intentionally underperform, there is considerable evidence that children can intentionally deny their knowledge or be successfully coached as to how to misrepresent themselves (Welsh, Bender, Whitman, Vasserman, & MacAllister, 2012). In situations where there is secondary gain to the child for underperforming (e.g., ADHD diagnosis with accompanying modifications and medications, ongoing legal activities in cases of TBI, underperforming in baseline pre-concussion testing) an additional source of potential error arises.

Research has demonstrated not only that children can feign cognitive impairment, but, more significantly, that it may be difficult for neuropsychologists to detect suboptimal effort by using clinical judgment alone (Faust, Hart, &

[Guilmette, 1988](#); [Faust, Hart, Guilmette, & Arkes, 1988](#)). In one of the first pediatric malingering studies conducted in the late 1980s, clinicians did not surpass chance in malingering identification. In another study conducted by the same researchers, three children between the ages of 9 and 12 years were asked to fake brain damage. Of the 42 clinicians who reviewed the test results, not one made that diagnosis. Clinical judgment was clearly not sufficient in judging valid effort. [Boone and Lu \(2003\)](#) have argued that validity tests routinely should be administered to any clients who would have incentive to underperform. Assessment of this form of testing error (i.e., intentionally underperforming) is accomplished through an individual's performance on testing (performance validity testing).

Performance validity testing (also referred to as malingering, response bias, and suboptimal effort) has been used both in research studies to exclude invalid performances and in forensic evaluations to infer the validity of other test performances in the battery ([Larrabee, 2012](#)). Typically, these types of tests are designed to have high levels (90%) of specificity (at the expense of high sensitivity) for the purpose of reducing false positives, that is, categorizing as invalid the performance of someone giving a validly impaired performance. Levels of atypical performance are validated by having patients with significant neurological disease (aphasia, traumatic brain injury, dementia), psychiatric disorders, anxiety, depression, emotional distress, and pain all perform at or near ceiling levels on the test ([Larrabee, 2012](#)).

There are three general approaches to performance validity testing: (1) forced-choice "stand-alone" measures; (2) nonforced choice stand-alone measures; and (3) embedded measures. Forced-choice measures have empirically derived cutoffs that have been developed to discriminate between examinees with genuine cognitive impairments and those simulating impairment. Each test has varying sensitivities and when compared some examinees will pass some measures while failing others. Thus, more than one forced-choice measure should be administered. Forced-choice measures are more time-consuming and are more easily recognized by an examinee. Also, most forced-choice measures are used to assess suboptimal effort on memory measures. Limiting performance validity testing to one cognitive domain may miss examinees trying to feign other impairments. Nonforced-choice measures are not as easily identified.

Embedded measures refers to standard neuropsychological tests that research has demonstrated are valid methods to identify suboptimal or disingenuous performance. The value of an embedded measure is that it is time efficient, and difficult for the examinee to identify. An example of this would be reliable digit span (RDS), which is calculated by summing the longest string of digits forward and backward that were provided without error across the two trials. Adult studies suggest that a cutoff of  $RDS \leq 7$  indicates poor effort (e.g., [Greiffenstein, Baker, & Gola, 1994](#); [Meyers & Volbrecht, 1998](#)). However, a study of children age 6–11 years found that 59% of participants "failed" using this criterion ([Blaskewitz, Merten, & Kathmann, 2008](#)). In a pediatric sample of patients with traumatic brain injuries, there was a significant difference in RDS scores between those who did and did not fail two neuropsychological measures; as a result, an RDS cutoff score of  $\leq 6$  was recommended as a threshold to detect effort in children ([Kirkwood, Hargrave, & Kirk, 2011](#)). Although the authors suggested that

RDS cutoff scores are likely associated with an elevated rate of false positives in children, they note that digit-span scores may have some utility as an embedded measure of effort in a relatively high-functioning pediatric sample. Again, in situations where the performance validity testing measures are positively correlated with age or IQ, the detection of true malingering is most reliable in older children with higher IQ scores.

Because clinical judgment has been shown to be ineffective in detecting examinee faking, multiple measures of effort should be incorporated into an evaluation. Also, observation alone cannot aid a clinician in determining whether someone is lying, exaggerating problems, or putting forth poor effort. One diagnostic area that has gained increasing attention in the neuropsychological assessment of performance validity testing in children and young adults is ADHD. There are multiple reasons why someone might feign an attention disorder. High school and college students are often hoping to have accommodations, especially extended time on standardized testing, and have easy access to psychostimulant medications, which are often misused by these individuals. The lack of a well-accepted neuropsychological gold standard for ADHD and the reliance on self-report and rating scales make the diagnosis even more difficult. [Boone and Lu \(2003\)](#) have suggested incorporating two performance validity tests into the neuropsychological test battery. To the extent that the two are not correlated, failure on two performance validity tests yields a false positive rate of .01. Again, these tests typically have much lower sensitivities so that not all suboptimal performances will be detected but if failure on two measures is the criterion, the likelihood of calling a valid performance, invalid will be very low. Finally, it is important to keep in mind that failure on two performance validity tests does not equate to malingering. Malingering requires intentional underperformance for the purpose of secondary gain. There are a variety of other reasons—emotional, emphasizing weakness, absence of motivation to perform well, etc.—that also elicit suboptimal performances.

### ***Clinical decision making***

Clinicians also are at serious risk of—unknowingly—generating diagnostic hypotheses based on nonpsychometric data and then seeking to validate these—unexamined—hypotheses by selecting psychometric data to fit. Clinicians, in any given instance, may be correct in their diagnoses. The problem is not only that they will not be correct in all of them, but also that they will not know in which instances they are correct and in which they are not. An additional concern is the fact that clinicians rely on the “objectivity” of the psychological tests, believing them to be a rigorous standard against which to establish diagnostic validity. Clinicians, however, may fail to appreciate that their use of the tests is by no means objective. This is no less of a potential and very worrisome problem in actuarially based assessment strategies as in more flexible and/or qualitative approaches ([Willis, 1986](#)).

The clinician is the primary analyzer of the behavioral information collected during the evaluation. In the clinical setting it is the clinician who brings an appreciation of the human condition and its vagaries to this encounter with the patient,

thus enriching in a uniquely human fashion the description of behavioral function provided by various measurement techniques. It is, however, the clinician's very humanness that makes him or her prone to error. The human mind is limited in its capacity to analyze information. Our attempts to circumvent this when confronted with complex cognitive tasks (as in the development of a diagnosis) lead to predictable types of decision-making biases. These are summarized in [Table 7.4](#).

**Table 7.4** Expectable biases in clinical decision-making

Decision-making bias	Nature of error
Limited capacity	<a href="#">Miller (1956)</a> argued that incremental improvements in decision-making accuracy occur until approximately 7 ( $\pm 2$ ) pieces of information have been collected. Beyond this, the capacity of the system is overloaded. Provision of additional information (beyond a few of the most valid indicators) is unlikely to increase predictive accuracy ( <a href="#">Oskamp, 1965</a> ) and may actually lead to a decrement in decision-making accuracy ( <a href="#">Golden, 1964; Wedding, 1983a, 1983b</a> ).
Simplification	When confronted with complex cognitive tasks, we typically resort to simplification of the information involved and are usually unaware of the manner in which these changes operate. Although we may believe that the exercise of complex pattern integration is what guides decision-making, it is the linear combination of data that has been shown to account for a range of diagnostic decisions made by, for example, radiologists ( <a href="#">Hoffman, Slovic, &amp; Rorer, 1968</a> ), psychiatrists ( <a href="#">Rorer, Hoffman, Dickman, &amp; Slovic, 1967</a> ), and psychologists ( <a href="#">Goldberg, 1968; Wiggins &amp; Hoffman, 1968</a> ). Indeed, <a href="#">Fisch, Hammond, and Joyce (1982)</a> demonstrated that psychiatrists' diagnoses of depression were fully accounted for by only one or two pieces of information despite their conviction that they were integrating many more data points into their decisions. Moreover, clinicians may believe that they rely upon a certain piece of information in making their decision, when analysis of their performance reveals that other information actually swayed their opinions ( <a href="#">Gauron &amp; Dickinson, 1966; Gauron &amp; Dickinson, 1969; Nisbett &amp; Wilson, 1977</a> ).
Use of heuristics	These simplifications can take the form of "rules of thumb" or intuitive heuristics that may be useful in certain applications, but lead to predictable types of decision-making biases.
Representativeness heuristic	When using this, people assess the probability of occurrence (i.e., graduate school major) by the degree to which the information is consistent with their preconceptions of the category ( <a href="#">Kahneman &amp; Tversky, 1972</a> ).

(Continued)

**Table 7.4** (Continued)

Decision-making bias	Nature of error
Availability heuristic	This is applied when people assess the probability of occurrence by the ease with which that occurrence can be remembered ( <a href="#">Tversky &amp; Kahneman, 1974</a> ). For example, they are influenced by factors such as recency (recent observance of a similar case) or the degree to which an outcome is memorable (unusual outcomes are salient simply because they are unexpected).
Anchoring and adjustment heuristic	These may shape probability judgments: the person starts from an initial value (anchor) which is usually the first piece of information examined and adjust their interpretation of additional data to conform.
Confirmatory bias	This leads people to emphasize information that is consistent with their hypotheses and to ignore information that would be contradictory to the hypothesis ( <a href="#">Nisbett &amp; Ross, 1980</a> ; <a href="#">Ross, Lepper, Strack, &amp; Steinmetz, 1977</a> ).
Covariation misestimation	As a result of these limitations on decision making, certain types of cognitive operations are difficult, if not impossible, to accomplish. For example, in order to determine whether the presence of a sign has a valid predictive relationship with an outcome (i.e., covariation), one must know: (1) the proportion of time that the sign is present when the condition is present and (2) the proportion of time that the sign is present when the condition is absent. Additionally, in order to know whether a sign has a useful diagnostic relationship to a given condition, one must know the base rate of the condition in the population of interest. This relationship is, however, extremely difficult to determine unless data are formally collected ( <a href="#">Arkes, 1981</a> ). Informal estimations of sign-outcome relationships are usually unreliable because preconceived notions can bias judgments of how variables covary (e.g., large eyes on human-figure drawings and suspicious personality types; <a href="#">Chapman &amp; Chapman, 1967</a> ).
Base rate neglect	A relationship between a sign and a condition may be valid and is certainly necessary to demonstrate the clinical utility of the sign; it is not, however, sufficient ( <a href="#">Faust &amp; Nurcombe, 1989</a> ). Information regarding the condition's base rate, that is, the prevalence of the condition within the population being examined ( <a href="#">Meehl &amp; Rosen, 1955</a> ), must also be considered. This is, however, not typically the case in clinical practice ( <a href="#">Kennedy, Willis, &amp; Faust, 1997</a> ).

We illustrate with two of these sources of potential error. First, the anchoring-and-adjustment bias can have significant impact on the nature of the information that is collected and/or considered important in different assessment strategies. A neurodevelopmental assessment model “anchors” on a review of systems and frames the clinical analysis in “brain” terms. A flexible battery approach is likely to anchor on the interview and history as the basis for selecting measures. A fixed battery approach anchors on the tests that constitute the battery. None of these strategies are free of the potential for “adjusting” subsequent data to match the framework provided by the initial data set; the way in which they adjust—and the controls necessary to accommodate potential for error—are likely to be different.

Second, under-utilization of base rate information, or base-rate neglect, is a common source of clinician error. Consider a test with 90% sensitivity (i.e., the proportion of individuals with a disorder who exhibit a sign) and 90% specificity (i.e., the proportion of individuals without a disorder who do not exhibit the sign). The parameter that the practicing clinician is most interested in is the test’s positive predictive power (PPP) that one actually has the disorder of interest (Ellwood, 1993). PPP is determined by the test’s sensitivity and specificity in the context of the base rate of the condition. Even with 90% sensitivity and specificity, if the base rate of the condition is relatively rare, the majority of individuals who exhibit that sign will not have the condition. Here, in a population of 1000 children, and a 4% base rate for ADHD, 40 children are expected to have ADHD. Using a test with 90% sensitivity and 90% specificity, however, only 27% of the children who receive an abnormal score on the test can be expected to actually have ADHD.

## Summary

The writings of Vygotsky, one of the most influential psychologists of the 20th century, are aligned with learning, educational psychology, and developmental and child neuropsychology. His theories on the zone of proximal development and the dynamic-systemic organization of psychological functions are central to our understanding of the child neuropsychological evaluation. Given our thesis that the cardinal feature of the child is development, it is clear that Vygotsky, with his intensive evaluation of the development of higher mental functions, is a founding parent of child neuropsychology.

Central purposes of the child neuropsychological evaluation include providing a diagnosis that might enable a child to receive services at school, helping to identify interventions that would benefit the child, clarifying future potential areas of risk, and enabling the adults working with the child to better understand him or her. The child neuropsychological evaluation differs from more common psychoeducational evaluations according to breadth, referral questions, training and experience of the clinician, and neurodevelopmental influences. The meaning of the child neuropsychological evaluation is derived from its ability to answer the question that prompted the evaluation. As the field has evolved common questions posed to

neuropsychologists have shifted from signs and locations of brain injuries to questions in the form of, "What is causing this child not to be at ease?" Here, ecologically valid assessments that support the capacity of the evaluation to confirm, validate, and interpret the referral concerns are important. There are a variety of neuropsychological assessment approaches, and we prefer those that are aligned with traditional standards of research methodology in seeking to disconfirm competing hypotheses.

Finally, as might be expected, there are a variety of sources of error that may obfuscate the validity of child neuropsychological evaluations. Here, we summarized issues of incremental validity, demographic characteristics of children, ecological validity, potential for suboptimal performance and malingering, and errors associated with clinical decision-making.

## References

- Akhutina, T. V. (2003). L. S. Vygotsky and A. R. Luria. Foundations of neuropsychology. *Journal of Russian & East European Psychology*, 41(3–4), 159–190.
- Anokhin, P. K. (1968). *Biology and neurophysiology of conditioned reflexes*. Moscow: Meditsina.
- Arkes, H. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology*, 49, 323–330.
- Benton, A. L., & Sivan, A. B. (2007). Clinical neuropsychology: A brief history. *Disease-a-Month*, 53, 142–147.
- Bergen, D., & Woodin, M. (2011). Neuropsychological development of newborns, infants, and toddlers (0–3 years old). In A. S. Davis (Ed.), *The handbook of pediatric neuropsychology* (pp. 15–30). New York: Springer Publishing Company.
- Bernstein, J. H., & Weiler, M. D. (2000). Pediatric neuropsychological assessment reexamined. In G. Goldstein (Ed.), *Handbook of psychological assessment*. New York: Pergamon Press, Inc.
- Blaskewitz, N., Merten, T., & Kathmann, N. (2008). Performance of children on symptom validity tests: TOMM, MSVT, and FIT. *Archives of Clinical Neuropsychology*, 23, 379–391.
- Bodrova, E. (2008). Make-believe play versus academic skills: A Vygotskian approach to today's dilemma of early childhood education. *European Early Childhood Education Research Journal*, 16, 357–369.
- Boone, K. B., & Lu, P. H. (2003). Noncredible cognitive performance in the context of severe brain injury. *The Clinical Neuropsychologist*, 17, 244–254.
- Boone, K., Victor, T., Wen, J., Razani, J., & Ponton, M. (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology*, 22, 355–365.
- Bricker, D., Cap, B., & Pretti-Frontczak, K. (2002). *Test for birth to three years and three to six years. Assessment, evaluation, and programming system for infants and children (AEPS)* (2nd ed.). Baltimore, MD: Paul H. Brookes Publishing Company.
- Brickman, A. M., Cabo, R., & Manly, J. J. (2006). Ethical issues in cross cultural neuropsychology. *Applied Neuropsychology*, 13, 91–100.

- Brinkman, J. J., Decker, S. L., & Dean, R. S. (2005). Assessing and understanding brain function through neuropsychologically based ability tests. In R. C. D'Amato, E. Fletcher-Janzen, & C. R. Reynolds (Eds.), *Handbook of school neuropsychology* (pp. 303–326). Hoboken, NJ: John Wiley & Sons.
- Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L., Dawson, D., ... Channon, S. (2006). The case for the development and use of “ecologically valid” measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, 12, 194–209.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press.
- Carroll, J. B. (2012). The three-stratum theory of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed, pp. 883–890). New York: Guilford Press.
- Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 72, 193–204.
- Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological skills. *Neuropsychology Review*, 113, 181–197.
- Clapper, T. C. (2015). Cooperative-based learning and the zone of proximal development. *Simulation and Gaming*, 46, 148–158.
- Cronbach, L. J., Rajarathnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberation or reliability theory. *The British Journal of Statistical Psychology*, 16, 137–163.
- Cubelli, R. (2005). The history of neuropsychology according to Norman Geschwind: Continuity and discontinuity in the development of science. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, 41, 271–274.
- Decker, S. L., Hale, J. B., & Flanagan, D. P. (2013). Professional practice issues in the assessment of intellectual development in children for educational applications. *Psychology in the Schools*, 50, 300–313.
- Ellwood, R. W. (1993). Clinical discriminations and neurapsychological tests: An appeal to Bayes’ theorem. *The Clinical Neuropsychologist*, 7, 224–233.
- Faust, D., Hart, K., & Guilmette, T. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, 56, 578–582.
- Faust, D., Hart, K., Guilmette, T., & Arkes, H. (1988). Neuropsychologists’ capacity to detect adolescent malingering. *Professional Psychology: Research and Practice*, 19, 508–515.
- Faust, D., & Nurcombe, B. (1989). Improving the accuracy of clinical judgment. *Psychiatry*, 52, 197–208.
- Fisch, H. U., Hammond, K. R., & Joyce, C. R. B. (1982). On evaluating the severity of depression: An experimental study of psychiatrists. *British Journal of Psychiatry*, 140, 378–383.
- Flanagan, D. P., Alfonso, V. C., & Ortiz, S. O. (2012). The cross-battery assessment approach: An overview, historical perspective, and current directions. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed, pp. 459–483). New York: Guilford Press.
- Fletcher-Janzen, E. (2005). The school neuropsychological evaluation. In R. C. D'Amato, E. Fletcher-Janzen, & C. R. Reynolds (Eds.), *Handbook of school neuropsychology* (pp. 172–212). Hoboken, NJ: John Wiley & Sons Inc.

- Fletcher, J. M., & Taylor, H. G. (1984). Neuropsychological approaches to children: Towards a developmental neuropsychology. *Journal of Clinical Neuropsychology*, 6, 39–56.
- Franzen, M. D., & Wilhelm, K. L. (1996). Conceptual foundations of ecological validity in neuropsychological assessment. In R. J. Sborfone (Ed.), *Ecological validity of neuropsychological testing* (pp. 91–112). Boca Raton, FL: St. Lucie Press.
- Gauron, E. F., & Dickinson, J. K. (1966). Diagnostic decision making in psychiatry: I. Information usage. *Archives of General Psychiatry*, 14, 225–232.
- Gauron, E. F., & Dickinson, J. K. (1969). The influence of seeing the patient first on diagnostic decision-making in psychiatry. *American Journal of Psychiatry*, 126, 199–205.
- Genesee, F. (2015). Myths about early childhood bilingualism. *Canadian Psychology*, 56, 6–15.
- Ghassemzadeh, H., Posner, M., & Rothbart, M. K. (2013). Contributions of Hebb and Vygotsky to an integrated science of mind. *Journal of the History of the Neurosciences*, 292–306.
- Glozman, J. (2013). *Developmental neuropsychology*. New York: Routledge.
- Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAT-II. *The Journal of Special Education*, 40, 103–114.
- Goldberg, L. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 23, 483–496.
- Golden, M. (1964). Some effects of combining psychological tests on clinical inferences. *Journal of Consulting Psychology*, 28, 440–446.
- Greiffenstein, M., Baker, W., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, 6, 218–224.
- Haagbloom, S. J., Warnick, R., Warnick, J., Jones, V. K., Yarbrough, G. L., Russell, T. M., ... Monte, E. (2002). The 100 most eminent psychologists of the 20th century. *Review of General Psychology*, 6, 139–152.
- Heilbronner, R. J., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S., & Conference Participants. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, 23, 1093–1129.
- Hoffman, P. J., Slovic, P., & Rorer, L. G. (1968). An analysis-of-variance model for the assessment of configural cue utilization in human judgment. *Psychological Bulletin*, 69, 338–349.
- Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *Woodcock–Johnson technical manual* (pp. 197–246). DLM: Allen, TX.
- Horn, J. L., & Blankson, N. (2012). Foundations for a better understanding of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed, pp. 73–98). New York: Guilford Press.
- Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. *Psychological Assessment*, 15, 443–445.
- Johnston, C., & Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents. *Psychological Assessment*, 15, 496–507.
- Jontak, J. (2013). Bilingual language development and language impairment in children. *Acta Neuropsychologica*, 13, 63–79.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.

- Keith, T. Z. (1994). Intelligence is important, intelligence is complex. *School Psychology Quarterly, 9*, 209–221.
- Keith, T. Z., & Reynolds, M. R. (2010). Cattell-Horn-Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47*, 635–650.
- Kennedy, M. L., Willis, W. G., & Faust, D. (1997). The base rate fallacy in school psychology. *Journal of Psychoeducational Assessment, 15*, 292–307.
- Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed). Fort Worth, TX: Holt, Rinehart & Winston.
- Kertesz, A. (1984). *Localization and neuroimaging in neuropsychology*. San Diego, CA: Academic Press.
- Kirkwood, M. W., Hargrave, D., & Kirk, J. W. (2011). The value of the WISC-IV Digit Span subtest in detecting noncredible performance during pediatric neuropsychological exam. *Archives of Clinical Neuropsychology, 26*, 377–384.
- Kirkwood, M. W., Kirk, J. W., Blaha, R. Z., & Wilson, P. (2010). Noncredible effort during pediatric neuropsychological exam: A case series and literature review. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence, 16*, 604–618.
- Knight, C., Alderman, N., & Burgess, P. W. (2002). Development of a simplified version of the muliple errands test for use in hospital settings. *Neuropsychological Rehabilitation, 12*, 231–255.
- Koziol, L. F., & Budding, D. E. (2011). Pediatric neuropsychological testing: Theoretical models of test selection and interpretation. In A. S. Davis (Ed.), *The Handbook of Pediatric Neuropsychology* (pp. 15–30). New York: Springer Publishing Company.
- Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*, 1–7.
- Manly, J. J. (2008). Critical issues in cultural neuropsychology: Profit from diversity. *Neuropsychology Review, 18*, 179–183.
- McGill, R. J. (2015). Interpretation of KABC-II scores: An evaluation of the incremental validity of Cattell–Horn–Carroll (CHC) factor scores in predicting achievement. *Psychological Assessment, 27*, 1417–1426.
- McGill, R. J., & Busse, R. T. (2015). Incremental validity of the WJ III COG: Limited predictive effects beyond the GIA-E. *School Psychology Quarterly, 30*, 353–365.
- Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4*, 268–273.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.
- Meyers, J. E., & Volbrecht, M. (1998). Validation of reliable digit span for detection of malingering. *Assessment, 5*, 303–307.
- Miller, D. C., & Maricle, D. E. (2012). The emergence of neuropsychological constructs into tests of intelligence and cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed, pp. 800–819). New York: Guilford Press.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.
- Mindt, M. R., Byrd, D., Saez, P., & Manly, J. J. (2010). Increasing culturally competent neuropsychological services for ethnic minority populations: A call to action. *The Clinical Neuropsychologist, 24*, 429–453.
- Murphy, E. A. (1979). *Probability in medicine*. Baltimore, MD: Johns Hopkins Press.

- Murphy, C., Scantlebury, K., & Milne, C. (2015). Using Vygotsky's zone of proximal development to propose and test an explanatory model for conceptualising coteaching in pre-service science teacher education. *Asia-Pacific Journal of Teacher Education*, 43(4), 281–295.
- Nici, J., & Reitan, R. M. (1986). Patterns of neuropsychological ability in brain-disordered versus normal children. *Journal of Consulting and Clinical Psychology*, 54, 542–545.
- Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.
- Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35, 250–256.
- Olson, K., & Jacobson, K. (2015). Cross-cultural considerations in pediatric neuropsychology: A review and call to attention. *Applied Neuropsychology: Child*, 4, 166–177.
- Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, 29, 261–265.
- Pennington, B. F. (1991). *Diagnosing learning disorders*. New York: Guilford Press.
- Price, K. J., Joscko, M., & Kerns, K. (2003). The ecological validity of pediatric neuropsychological tests of attention. *The Clinical Neuropsychologist*, 17, 170–181.
- Rabinowitz, J. (1994). Guide to identifying and correcting decision-making errors in mental disability practice. *Bulletin of the American Academy of Psychiatry and Law*, 22, 561–575.
- Reitan, R. M. (1989). A note regarding some aspects of the history of clinical neuropsychology. *Archives of Clinical Neuropsychology*, 385–391.
- Rorer, L. G., Hoffman, H. D., Dickman, H. D., & Slovic, P. (1967). Configural judgments revealed (summary). *Proceedings of the 75th Annual Convention of the American Psychological Association*, 2, 195–196.
- Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.
- Ross, L. D., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality & Social Psychology*, 35, 817–829.
- Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition*, 52, 326–333.
- Salekin, R. T., Kubak, F. A., & Lee, Z. (2008). *Deception in children and adolescents: Clinical assessment of malingering and deception* (3rd ed, pp. 343–364). New York: Guilford Press.
- Schneider, W. J., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed, pp. 99–144). New York: Guilford Press.
- Silver, C. H., Blackburn, L. B., Arffa, S., Barth, J. T., Bush, S. S., Koffler, S. P., ... Elliott, R. W. (2006). The importance of neuropsychological assessment for the evaluation of childhood learning disorders: NAN policy and planning committee. *Archives of Clinical Neuropsychology*, 21(7), 741–744.
- Stuss, D. T., Binns, M. A., Murphy, K. J., & Alexander, M. P. (2002). Dissociations within the anterior attentional system: Effects of task complexity and irrelevant information on reaction time speed and accuracy. *Neuropsychology*, 16, 500–513.
- Tramontana, M. B. (1983). Neuropsychological evaluation of children and adolescents with psychopathological disorders. In P. J. Golden, & E. D. Vincente (Eds.), *Foundations of Clinical Neuropsychology* (pp. 309–340). New York: Plenum Press.
- Tramontana, M. B., & Hooper, S. R. (1997). Neuropsychology of child psychopathology. In C. R. Reynolds, & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 120–137). New York: Plenum Press.

- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 183*, 1124–1131.
- United States Census. (2013). Retrieved from <https://www-census-gov.uri.idm.oclc.org/data-tables/2013/demo/2009-2013-lang-tables.html>.
- United States Census. (2015). Retrieved from <https://www-census-gov.uri.idm.oclc.org/news-room/press-releases/2015/cb15-185.html>.
- van Der Veer, R. (2015). Vygotsky, the theater critic: 1922–3. *History of the Human Sciences, 28*, 103–110.
- van Der Veer, R., & Yasnitsky, A. (2011). Vygotsky in English: What still needs to be done. *Integrative Psychological & Behavioral Science, 45*, 475–493.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Oxford, England: Harvard University Press.
- Vygotsky, L. S. (1987). *The collected works of L. S. Vygotsky: Vol. 1. Problems of general psychology*. New York: Plenum Press.
- Vygotsky, L. S. (1997). *The collected works of L. S. Vygotsky: Vol. 3. Problems of the theory and history of psychology*. New York: Plenum Press.
- Wechsler, D. (2014). *Wechsler intelligence Scale for Children: Administration and scoring manual* (5th ed). Bloomington, MN: PsychCorp.
- Wedding, D. (1983a). Clinical and statistical prediction in neuropsychology. *Clinical Neuropsychology, 5*, 49–55.
- Wedding, D. (1983b). Comparison of statistical and actuarial models for predicting lateralization of brain damage. *Clinical Neuropsychology, 5*, 15–20.
- Welsh, A. J., Bender, A., Whitman, L. A., Vasserman, M., & MacAllister, W. S. (2012). Clinical utility of Reliable Digit Span in assessing effort in children and adolescents with epilepsy. *Archives of Clinical Neuropsychology, 27*, 735–741.
- Wiggins, N., & Hoffman, P. J. (1968). Three models of clinical judgment. *Journal of Abnormal Psychology, 73*, 70–77.
- Wilkening, G. N. (1989). Techniques of localization in child neuropsychology. In C. R. Reynolds, & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 291–309). New York: Penum Press.
- Willis, W. G. (2017). Intelligence tests. In E. Flannery-Schroeder, & D. Friedman-Wheeler (Eds.), *The SAGE encyclopedia of abnormal and clinical psychology* (pp. 1856–1859). Los Angeles, CA: Sage Publications.
- Willis, W. G. (1986). Actuarial and clinical approaches to neuropsychological diagnosis: Applied considerations. In J. E. Obrzut, & G. W. Hynd (Eds.), *Child neuropsychology: Vol. 2. Clinical practice*. Orlando, FL: Academic Press.
- Zonzi, A., Barkham, M., Hardy, G., Llewelyn, S., Stiles, W., & Leiman, M. (2014). Zone of proximal development (ZPD) as an ability to play in psychotherapy: A theory-building case study of very brief therapy. *Psychology and Psychotherapy: Theory, Research, and Practice, 87*, 447–464.

## Further reading

- Akhutina, T. V., & Pylaeva, N. M. (2011). L. Vygotsky, A. Luria and developmental neuropsychology. *Psychology in Russia: State of the Art, 155–175*.

- Bodrova, E., Leong, D. J., & Akhutina, T. V. (2011). When everything new is well-forgotten old: Vygotsky/Luria insights in the development of executive functions. *New Directions for Child and Adolescent Development*, 133, 11–28.
- Ghassemzadeh, H. (2009). Vygotsky in Iran: A personal account. *Journal of Cultural-Historical Psychology*, 4, 7–9.
- Goldstein, S., & Reynolds, C. R. (2011). Introduction. In S. Goldstein, & C. R. Reynolds (Eds.), *Handbook of neurodevelopmental and genetic disorders in children* (pp. 3–8). New York: Guilford Press.
- Reitan, R. M. (1959). The comparative effects of brain damage on the halstead impairment index and the Wechsler–Bellevue scale. *Journal of Clinical Psychology*, 15, 281–285.
- Sbordone, R. J. (1996). Ecological validity: Some critical issues for the neuropsychologist. In R. J. Sbordone (Ed.), *Ecological validity of neuropsychological testing* (pp. 15–41). Delray Beach, FL: GR Press/St. Lucie Press.

# Adult comprehensive neuropsychological assessment

8

Gerald Goldstein<sup>1</sup>, Daniel N. Allen<sup>2</sup> and John DeLuca<sup>3</sup>

<sup>1</sup>VA Pittsburgh Healthcare System, Pittsburgh, PA, United States, <sup>2</sup>Department of Psychology, University of Nevada, Las Vegas, NV, United States, <sup>3</sup>Kessler Foundation Research Center, West Orange, NJ, United States

## Introduction

This chapter provides a general introduction to the field of comprehensive neuropsychological assessment and deals specifically with the extensive standard test batteries used with adults. Neuropsychological assessment is a relatively new term that has essentially replaced the older terms “testing for brain damage” or “testing for organicity.” [Lezak \(1995\)](#) indicated that these procedures are used for three purposes: diagnosis, provision of information important for patient care, and research. A significant component of the patient care function is rehabilitation planning and monitoring ([Goldstein, 1978](#); [Goldstein & Beers, 1998](#); [Goldstein & Ruthven, 1983](#); [Meier, Benton, & Diller, 1987](#)). The focus of neuropsychological assessment has traditionally been on the brain-damaged patient, but there have been major extensions of the field to psychiatric disorders ([Goldstein, 1986a, 1986b, 1991](#); [Goldstein, Shemansky, & Allen, 2005](#); [Yozawitz, 1986](#)), functioning of nonbrain-damaged individuals with medical disorders ([Ryan, 1998](#)) and normal aging ([Goldstein & Shelly, 1975](#); [Nussbaum, 1997](#)).

Perhaps the best definition of a neuropsychological test has been offered by Ralph Reitan, who described it as a test that is sensitive to the condition of the brain. If performance on a test changes with a change in brain function then the test is a neuropsychological test. However, it should be pointed out that the comprehensive neuropsychological test batteries should not only contain neuropsychological tests, they should also contain some tests that are generally insensitive to brain dysfunction, primarily because such tests are often useful for providing a baseline against which the extent of impairment associated with acquired brain damage can be measured. Most neuropsychological assessment methods are formal tests, but some work has been done with rating scales and self-report measures. Neuropsychological assessment is rarely conducted through a structured interview outside of a test situation.

A comprehensive neuropsychological test battery is a procedure that assesses all of the major functional areas generally affected by structural brain damage. We use the term ideally because none of the standard, commonly available procedures entirely achieve full comprehensiveness. Some observers have described the

comprehensive procedures as screening batteries, because feasibility and time constraints generally require a sacrifice of detailed investigations of specific areas in order to achieve comprehensiveness. In specialized neuropsychological assessment we consider what a clinical neuropsychologist does when asked to explore a particular area in detail rather than do a comprehensive evaluation. While the term screening may be justifiable in certain respects, extensive standard batteries in common use should not be grouped with the brief, paper-and-pencil screening tests used in many clinical and industrial settings. That is, they do not simply screen for presence or absence of brain damage, but also evaluate a number of functional areas that may be affected by brain damage. Since brain damage most radically affects cognitive processes, most neuropsychological tests assess various areas of cognition, but perception and motor skills are also frequently evaluated. Thus, neuropsychological tests are generally thought of as assessment instruments for a variety of cognitive, perceptual, and motor skills. That is not to say that brain damage does not affect other aspects of the personality, but traditionally the standard neuropsychological tests do not typically assess these other areas. Perhaps the most important reason for this preference is that cognitive tests have proven to be the most diagnostic ones. While personality changes may occur with a wide variety of psychiatric, general medical and neurological conditions, cognitive changes appear to occur most dramatically in individuals with structural brain damage. Numerous attempts have been made to classify the functional areas typically affected by brain damage, but the scheme proposed in what follows is a reasonably representative one.

Perhaps the most ubiquitous change is general intellectual impairment. Following brain damage, the patient is not as bright as he or she was before. Problems are solved less effectively, goal-directed behavior becomes less well organized, and there is impairment of a number of specific skills, such as solving arithmetic problems or interpreting proverbs. Numerous attempts have been made to epitomize this generalized loss, perhaps the most effective one, articulated some time ago, being [Goldstein and Scheerer's \(1941\)](#) concept of impairment of the abstract attitude. The abstract attitude is a phenomenological concept having to do with the way in which the individual perceives the world. Some consequences of its impairment involve failure to form concepts or to generalize from individual events, failure to plan ahead ideationally, and inability to transcend the immediate stimulus situation. While the loss is a general one involving many aspects of the individual's life, it is best observed in a testing setting where the patient is presented with a novel situation in which some problem must be solved. Typically these tests involve abstraction or concept formation, and the patient is asked to sort or categorize in some way. The [Goldstein—Scheerer tests \(1941\)](#), perhaps the first neuropsychological battery, consist largely of sorting tests, but also provide the patient with other types of novel problem-solving tasks.

Probably the next most common manifestation of structural brain damage is impairment of memory. Sometimes memory impairment is associated with general intellectual impairment, sometimes it exists independently, and sometimes it is seen as an early sign of a progressive illness that eventually impairs a number of abilities other than memory. In most, but not all cases, recent memory is more impaired

than remote memory. That is, the patient may recall ‘his or her early life in great detail, but may be unable to recall what happened during the previous day. Often, so-called primary memory is also relatively well preserved. That is, the patient may be able to immediately repeat back what was just presented, such as a few words or a series of digits, but will not retain new information over a more extended period of time, particularly after intervening events have occurred. In recent years, our capacity to examine memory has benefited from a great deal of research involving the various amnesia syndromes (e.g., Baddeley, Wilson, & Watts, 1995; Butters & Cermak, 1980), and we have been quite aware for some time that not all brain-damaged patients experience the same kind of memory disorder (Butters, 1983). It generally requires a detailed assessment to specifically identify the various types of memory disorder, and the comprehensive batteries we will be discussing here generally can only detect the presence of a memory disorder and provide an index of its severity.

Loss of speed in performing skilled activities is an extremely common symptom of brain damage. Generally, this loss is described in terms of impaired psychomotor speed or perceptual-motor coordination. While its basis is sometimes reduction of pure motor speed, in many instances pure speed is preserved in the presence of substantial impairment on tasks involving speed of some mental operation or coordination of skilled movement with perceptual input. Thus, the patient may do well on a simple motor task such as finger tapping, but poorly on a task in which movement must be coordinated with visual input, such as a cancellation or substitution task. Tasks of this latter type are commonly performed slowly and laboriously by many kinds of brain-damaged patients. Aside from slowness, there may be other disturbances of purposive movement that go under the general heading of apraxia. Apraxia may be manifested as simple clumsiness or awkwardness, an inability to carry out goal-directed movement sequences as would be involved in such functional activities as dressing, or as an inability to use movement ideationally as in producing gestures or performing pretended movements. While apraxia in one of its pure forms is a relatively rare condition, impairment of psychomotor speed is quite common and seen in a variety of conditions.

A set of abilities that bridge movement and perception may be evaluated by tasks in which the patient must produce some form of construction or copy a figure from a model. Among the first tests used to test brain-damaged patients was the Bender-Gestalt (Bender, 1938) a procedure in which the patient must copy a series of figures devised by Wertheimer (1923) to study perception of visual gestalten. It was found that many patients had difficulty copying these figures, although they apparently perceived them normally. These difficulties manifested themselves in reasonably characteristic ways, including various forms of distortion, rotation of the figure, simplification, or primitivation and perseveration. The copying task has continued to be used by neuropsychologists, either in the form of the Bender-Gestalt or a variety of other procedures. Variations of the copying task procedure have involved having the patient draw the figure from memory (Benton, 1963; Rey, 1941), from a verbal command, for example, “Draw a Circle” (Luria, 1973), or copy a figure that is embedded in an interfering background pattern (Canter, 1970). The Rey-Osterrieth Complex Figure Test involving copying of a

very detailed figure is probably the most commonly used of these tests at present ([Osterrieth, 1944](#)).

Related to the copying task is the constructional task, in which the patient must produce a three-dimensional construction from a model. The most popular test for this purpose is the Kohs Blocks or Block Design subtest of the Wechsler scales ([Wechsler, 1997a, 1997b](#)). While in the timed versions of these procedures the patient may simply fail the task by virtue of running out of time, at least some brain-damaged patients make errors on these procedures comparable to what is seen on the copying tasks. With regard to block-design type tasks, the errors might involve breaking the contour of the model or incorrectly reproducing the internal structure of the pattern ([Kaplan, 1979](#)). Thus, a constructional deficit may not be primarily associated with reduction in psychomotor speed, but rather the inability to build configurations in three-dimensional space. Often, the ability is referred to as visual–spatial skill.

Visual–spatial skills also form a bridge with visual perception. When one attempts to analyze the basis for a patient’s difficulty with a constructional task, the task demands may be broken down into movement, visual, and integrative components. Often, the patient has no remarkable impairment of purposive, skilled movement and can recognize the figure. If it is nameable, the patient can tell you what it is or if it is not, it can be correctly identified on a recognition task. However, the figure cannot be accurately copied. While the difficulty may be with the integration between the visual percept and the movement, it has also been found that patients with constructional difficulties, and indeed patients with brain damage in general, frequently have difficulties with complex visual perception. For example, they do poorly at embedded figures tasks ([Teuber, Battersby, & Bender, 1951](#)) or at tasks in which a figure is made difficult to recognize through displaying it in some unusual manner, such as overlapping it with other figures ([Golden, 1981](#)) or presenting it in some incomplete or ambiguous form ([Mooney, 1957; Warrington & James, 1991](#)). Some brain-damaged patients also have difficulty when the visual task is made increasingly complex through adding elements in the visual field. Thus, the patient may identify a single element, but not two. When two stimuli are presented simultaneously, the characteristic error is that the patient reports only seeing one. The phenomenon is known as extinction or neglect ([Bender, 1952](#)).

Many brain-damaged patients also have deficits in the areas of auditory and tactile perception. Sometimes, the auditory impairment is such that the patient can hear, but sounds cannot be recognized or interpreted. The general condition is known as agnosia and can actually occur in the visual, auditory, or tactile modalities. Agnosia has been defined as “perception without meaning,” implying the intactness of the primary sense modality but loss of the ability to comprehend the incoming information. Auditory agnosia is a relatively rare condition, but there are many disturbances of auditory perception that are commonly seen among brain-damaged patients. Auditory neglect can exist, and it is comparable to visual neglect; sounds to either ear may be perceived normally, but when a sound is presented to each ear simultaneously, only one of them may be perceived. There are a number of auditory verbal problems that we will get to when we discuss language. Auditory

attentional deficits are common and may be identified by presenting complex auditory stimuli, such as rhythmic patterns, which the patient must recognize or reproduce immediately after presentation. A variety of normal and abnormal phenomena may be demonstrated using a procedure called dichotic listening (Kimura, 1961). It involves presenting two different auditory stimuli simultaneously to each ear. The subject wears ear-phones, and the stimuli are presented using stereophonic tape. Higher level tactile deficits generally involve a disability with regard to identifying symbols or objects by touch. Tactile neglect may be demonstrated by touching the patient over a series of trials with single and double stimuli, and tactile recognition deficits may be assessed by asking the patient to name objects placed in his or her hand or to identify numbers or letters written on the surface of the skin. It is particularly difficult to separate primary sensory functions from higher cognitive processes in the tactile modality and neuropsychologists may perform rather detailed sensory examinations of the hands, involving such matters as light touch thresholds, two-point discrimination, point localization, and the ability to distinguish between sharp and dull tactile stimuli (Golden, Purisch, & Hammeke, 1985; Semmes, Weinstein, Ghent, & Teuber, 1960).

The neuropsychological assessment of speech and language has in some respects become a separate discipline involving neuropsychologists, neurologists, and speech and language pathologists. There is an extensive interdisciplinary literature in the area (Albert, Goodglass, Helm, Ruben, & Alexander, 1981; Benson & Ardila, 1996), and several journals that deal almost exclusively with the relationships between impaired or normal brain function and language (e.g., *Brain and Language*). Aphasia is the general term used to denote impairment of language abilities as a result of structural brain damage, but not all brain-damaged patients with communicative difficulties have aphasia. While aphasia is a general term covering numerous subcategories, it is now rather specifically defined as an impairment of communicative ability associated with focal damage to the left hemisphere in most people. Stroke is probably the most common cause of aphasia.

Historically, there have been numerous attempts to categorize the subtypes of aphasia (Goodglass, 1983), but in functional terms, the aphasias involve a rather dramatic impairment of the capacity to speak, to understand the speech of others, to find the names for common subjects, to read (alexia), write (agraphia), calculate (acalculia), or to use or comprehend gestures. However, a clinically useful assessment of these functional disorders must go into their specific characteristics. For example, when we say the patient has lost the ability to speak, we may mean that he or she has become mute or can only produce a few utterances in a halting, labored way. On the other hand, we may mean that the patient can produce words fluently but the words and sentences being uttered make no sense. When it is said that the patient does not understand language that may mean that spoken but not written language is understood, or it may mean that all modalities of comprehension are impaired. Thus, there are several aphasic syndromes, and it is the specific syndrome that generally must be identified in order to provide some correlation with the underlying localization of the brain damage and to make rational treatment plans. We may note that the standard comprehensive neuropsychological test

batteries do not include extensive aphasia examinations. There are several such examinations available, such as the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983) and the Western Aphasia Battery (Kertesz, 1979). Even though they may be used in conjunction with a neuropsychological assessment battery, they are rather lengthy procedures in themselves and require special expertise to administer and interpret.

For various reasons, it is often useful to assess attention as part of the neuropsychological examination. Sometimes an attention deficit is a cardinal symptom of the disorder, but even if it isn't, the patient's level of attention may influence performance on tests of essentially all of the functional areas we have been discussing. A discussion of attention may be aided by invoking a distinction between wide-aperture and narrow-aperture attention (Kinsbourne, 1980). Wide-aperture attention has to do with the individual's capacity to attend to an array of stimuli at the same time. Attention may be so narrowly focused that the total picture is not appreciated. Tests for neglect may in fact be assessing wide-aperture attention; narrow-aperture attention has to do with the capacity to sustain attention to small details. Thus, it can be assessed by vigilance tasks or tests like the Picture Completion subtest of the Wechsler scales. Brain-damaged patients may manifest attentional deficits of either type. They may fail to attend to a portion of their perceptual environment, or they may be unable to maintain sufficient concentration to successfully complete tasks requiring sustained occupation with details. Individuals with attentional deficits are often described as distractible or impulsive, and in fact, many brain-damaged patients may be accurately characterized by those terms. Thus, the assessment of presence and degree of attention deficit is often a highly clinically relevant activity. Mirsky, Anthony, Duncan, Ahearn, and Kellam (1991) have proposed a useful division, based on a factor-analytic study of attentional tasks, dividing them into tests that evaluate encoding, sustaining concentration, focusing, and shifting attention from one aspect of a task to another.

In summary, comprehensive neuropsychological assessment typically involves the functional areas of general intellectual capacity, memory, speed and accuracy of psychomotor activity, visual-spatial skills, visual, auditory, and tactile perception, language, and attention. Thus, a comprehensive neuropsychological assessment may be defined as a procedure that at least surveys all of these areas. In practical terms, a survey is all that is feasible if the intent of the assessment is to evaluate all areas. It is obviously generally not feasible to do an in-depth assessment of each of these areas in every patient, nor is it usually necessary to do so.

## **Special problems in the construction and standardization of neuropsychological test batteries**

It will be assumed here that neuropsychological tests share the same standardization requirements as all psychological tests. That is, there is the need for appropriate quantification, norms, and related test-construction considerations, as well as the

need to deal with issues related to validity and reliability. The chapter in this book by Reynolds and Livingston describe the matters in detail. However, there are some special considerations regarding neuropsychological tests, and we will turn our attention to them here.

### ***Practical concerns in test construction***

Neuropsychological test batteries must of necessity be administered to brain-damaged patients, many of whom may have severe physical disability, cognitive impairment, or a combination of the two. Thus, stimulus and response characteristics of the tests themselves, as well as the stimulus characteristics of the test instructions, become exceedingly important considerations. Neuropsychological test material should, in general, be constructed with salient stimuli that the patient can readily see or hear and understand. Material to be read should not require high levels of literacy, nor should grammatical structures be unduly complex.

With regard to test instruction, the potential for multimodal instruction-giving should ideally be available. If the patient cannot see or read it should be possible to say the instructions, without jeopardizing one's opportunity to use established test norms. The opportunity should be available to repeat and paraphrase instructions until it is clear that they are understood. It is of crucial importance in neuropsychological assessment that the examiner achieves maximum assurance that a test was failed because the patient could not perform the task being assessed, not because the test instructions were not understood. This consideration is of particular importance for the patient with aphasia, who may have a profound impairment of language comprehension. With regard to response parameters, efforts should be made to assure that the test response modality is within the patient's repertoire.

In neuropsychological assessment, it is often not failure to perform some specific task that is diagnostic, but failure to perform some component of a series of tasks in the presence of intact function in other areas. As an example, failure to read a passage is not specifically diagnostic, since the inability to read may be associated with a variety of cognitive, perceptual, and learning difficulties. However, failure to be able to transfer a grapheme or a written symbol to a phoneme or sound in the presence of other manifestations of literacy could be quite diagnostic. Individuals with this type of deficit may be able to "sight-read" or recognize words as perceptual patterns, but when asked to read multisyllabic, unfamiliar words, they are unable to break the word down into phonemes and sound it out. In perhaps its most elegant form, neuropsychological assessment can produce what is called a double dissociation ([Teuber, 1959](#)); a task consistently failed by patients with a particular type of brain disorder accompanied by an equally difficult corresponding task that is consistently passed, and the reverse in the case of patients with some other form of brain disorder. Ideally, then, neuropsychological assessment aims at detailed-as-possible specification of what functional deficits exist in a manner that allows for mapping of these deficits onto known systems in the brain. There are several methods of achieving this goal, and not all neuropsychologists agree with regard to the most productive route. In general, some prefer to examine patients in what may be

described as a linear manner, with a series of interlocking component abilities, while others prefer using more complex tasks in the form of standard, extensive batteries and interpretation through examination of performance configurations. The linear approach is best exemplified in the work of Luria and various collaborators (Luria, 1973), while the configurational approach is seen originally in the work of Ward Halstead (Halstead, 1947) and Ralph Reitan (Reitan & Wolfson, 1993) and their many collaborators. In either case, however, the aim of the assessment is largely that of determining the pattern of the patient's preserved and impaired functions and inferring from this pattern what the nature might be of the disturbed brain function. The difficulty with using complex tasks to achieve that end is that such tasks are really only of neuropsychological interest if they can be analyzed by one of the two methods described here.

### ***Issues related to validity and reliability***

Neuropsychological assessment has the advantage of being in an area where the potential for development of highly sophisticated validation criteria has been very much realized in recent years and will surely continue to achieve even fuller realization in the future. We will begin our discussion with this consideration, and so we will first be occupied with the matters of concurrent and predictive validity. A major review of validation studies was accomplished by Kjove (1974) and updated by Boll (1981). A recent review was done by Reed and Reed (1997). Reitan and Wolfson (1993) have written an entire volume on the Halstead-Reitan battery (HRB) which contains a brief review of pertinent research findings in addition to extensive descriptions of the tests themselves and case materials. These reviews essentially only covered the Wechsler scales and the HRB, but there are several reviews of the work with the Luria-Nebraska Neuropsychological Battery (LNNB) as well (e.g., Moses & Purisch, 1997).

We will not deal with the content of those reviews at this point, but rather focus on the methodological problems involved in establishing concurrent or predictive validity of neuropsychological tests. With regard to concurrent validity, the criterion used in most cases is the objective identification of some central-nervous-system lesion arrived at independently of the neuropsychological test results. Therefore, validation is generally provided by neurologists or neurosurgeons. Identification of lesions of the brain is particularly problematic because, unlike many organs of the body, the brain cannot usually be visualized directly in the living individual. The major exceptions occur when the patient undergoes brain surgery or receives the rarely used procedure of brain biopsy. In the absence of these procedures, validation is dependent upon autopsy data or the various brain-imaging techniques. Autopsy data are not always entirely usable for validation purposes, in that numerous changes may have taken place in the patient's brain between time of testing and time of examination of the brain. Of the various imaging techniques, magnetic resonance imaging (MRI) is currently the most commonly used one. Cooperation among neuroradiologists, neurologists, and neuropsychologists has already led to the accomplishment of several important studies correlating quantitative magnetic

resonance data with neuropsychological test results (e.g., [Minshew, Goldstein, Dombrowski, Panchalingam, & Petlegrew, 1993](#)). Beyond MRI, however, we can see even more sensitive indicators, including measures of cerebral metabolism such as the PET scan (Positron Emission Tomography), and functional MRI. Recently, more generally available and even more sensitive measures of cerebral metabolism have appeared including more recent generations of the PET scan, allowing for greatly improved resolution, SPECT (Single Photon Emission Computerized Tomography), which allows for studying brain metabolism in settings in which a cyclotron is not available, and the evolving methods of diffusion tensor imaging (DTI) and magnetic resonance spectroscopy (MRS). These exciting new developments in brain imaging and observation of brain function will surely provide increasingly definitive criteria for neuropsychological hypothesis testing and assessment methods.

Within neuropsychological assessment, there has been a progression regarding the relationship between level of inference and criterion. Early studies in the field as well as the development of new assessment batteries generally addressed themselves to the matter of simple presence or absence of structural brain damage. Thus, the first question raised had to do with the accuracy with which an assessment procedure could discriminate between brain-damaged and nonbrain-damaged patients, as independently classified by the criterion procedure. In the early studies, the criterion utilized was generally clinical diagnosis, perhaps included in some cases with neurosurgical data or some laboratory procedure such as a skull X-ray or an EEG. It soon became apparent, however, that many neuropsychological tests were performed at abnormal levels, not only by brain-damaged patients, but by patients with several of the functional psychiatric disorders. Since many neuropsychologists worked in neuropsychiatric rather than general medical settings, this matter became particularly problematic. Great efforts were then made to find tests that could discriminate between brain-damaged and psychiatric patients or, as sometimes put, between "functional" and "organic" conditions. There have been several early reviews of this research ([Goldstein, 1978](#); [Heaton, Baade, & Johnson, 1978](#); [Heaton & Crowley, 1981](#); [Malec, 1978](#)), all of which were critical of the early work in this field in light of current knowledge about several of the functional psychiatric disorders. The chronic schizophrenia patient was particularly problematic, since such patients often performed on neuropsychological tests in a manner indistinguishable from the performance of patients with generalized structural brain damage. By now, this whole issue has been largely reformulated in terms of looking at the neuropsychological aspects of many of the functional psychiatric disorders (e.g., [Goldstein, 1991](#); [Henn & Nasrallah, 1982](#)), largely under the influence of the newer biological approaches to psychopathology.

Neuropsychologists working in neurological and neurosurgical settings were becoming increasingly interested in validating their procedures against more refined criteria, notably in the direction of localization of brain function. The question was no longer only whether a lesion was present or absent, but if present, whether or not the tests could predict its location. Major basic research regarding this matter was conducted by Teuber and various collaborators over a span of many years

(Teuber, 1959). This group had access to a large number of veterans who had sustained open head injuries during World War II and the Korean conflict. Because the extent and site of their injuries were exceptionally well documented by neurosurgical and radiological data, and the lesions were reasonably well localized, these individuals were used productively in a long series of studies in which attempts were made to relate both site of lesion and concomitant neurological defects to performance on an extensive series of neuropsychological procedures ranging from measures of basic sensory functions (Semmes et al., 1960) to complex cognitive skills (Teuber & Weinstein, 1954). Similar work with brain-wounded individuals was accomplished by Freda Newcombe and collaborators at Oxford (Newcombe, 1969). These groups tended to concentrate on the major lobes of the brain (frontal, temporal, parietal, and occipital), and would, for example produce contrasts between the performances of patients with frontal and occipital lesions on some particular test or test series (e.g., Teuber, 1964).

In another setting, but at about the same time as the Teuber group was beginning its work, Ward Halstead and collaborators conducted a large-scale neuropsychologically oriented study of frontal lobe function (Halstead, 1947). Ralph M. Reitan, who was Halstead's student, adopted several of his procedures, supplemented them, and developed a battery of tests that were extensively utilized in localization studies. Reitan's early work in the area of localization was concerned with differences between the two cerebral hemispheres more than with regional localization (Reitan, 1955). The now well-known Wechsler–Bellevue studies of brain lesion lateralization [see review in Reitan (1966)] represented some of the beginnings of this work. The extensive work of Roger Sperry and various collaborators (Sperry, Gazzaniga, & Bogen, 1969) with patients who had undergone cerebral commissurotomy also contributed greatly to validation of neuropsychological tests with regard to the matter of differences between the hemispheres; particularly the functional asymmetries or cognitive differences. Since the discoveries regarding the major roles of subcortical structures in the mediation of various behaviors (Cummings, 1990), neuropsychologists have also been studying the relationships between test performance and lesions in such structures and structure complexes as the limbic system (Scoville & Milner, 1957) and the basal ganglia (Butters, 1983).

The search for validity criteria has become increasingly precise with recent advances in the neurosciences as well as increasing opportunities to collect test data from various patient groups. One major conceptualization largely attributable to Reitan and his coworkers is that localization does not always operate independently with regard to determination of behavioral change, but interacts with type of lesion or the specific process that produced the brain damage. The first report regarding this matter related to differences in performance between patients with recently acquired lateralized brain damage and those who sustained lateralized brain damage at some time in the remote past (Fitzhugh, Fitzhugh, & Reitan, 1961; Fitzhugh, Fitzhugh, & Reitan, 1962). Patients with acute lesions were found to perform differently on tests than patients with chronic lesions. It soon became apparent, through an extremely large number of studies (e.g., Goldstein, Nussbaum, & Beers, 1998) that there are many forms of type–locus interactions, and that level and pattern of

performance on neuropsychological tests may vary greatly with the particular nature of the brain disorder. This development paralleled such advances in the neurosciences as the discovery of neurotransmitters and the relationship between neurochemical abnormalities and a number of the neurological disorders that historically had been of unknown etiology. We therefore have the beginnings of the development of certain neurochemical validating criteria (Deutsch & Davis, 1983; Freedman, 1990). There has also been increasing evidence for a genetic basis for several mental and neurological disorders. The gene for Huntington's disease has been discovered, and there is growing evidence for a significant genetic factor contributing to the acquisition of certain forms of alcoholism (Steinhauer, Hill, & Zubin, 1987). In general, the concurrent validity studies have been quite satisfactory, and many neuropsychological test procedures have been shown to be accurate indicators of many parameters of brain dysfunction.

A persistent problem in the past has been the possible tendency of neuropsychological tests to be more sensitive than the criterion measures. In fact, a study by Filskov and Goldstein (1974) demonstrated that neuropsychological tests may predict diagnosis more accurately than many of the individual neurodiagnostic procedures commonly used in assessment of neurological and neurosurgical patients (e.g., skull X-ray). It would appear that with the advent of the MRI scan and the even more advanced brain-imaging procedures this problem has diminished. A related problem involves the establishment of the most accurate and reliable external criterion. We have always taken the position (Goldstein & Shelly, 1982; Russell, Neuringer, & Goldstein, 1970) that no one method can be superior in all cases, and that the best criterion is generally the final medical opinion based on a comprehensive but pertinent evaluation, excluding, of course, behavioral data. In some cases, for example, the MRI scan may be relatively noncontributory, but there may be definitive laboratory findings based on examination of blood or cerebral spinal fluid. In some cases (e.g., Huntington's disease) the family history and genetic studies may be the most crucial part of the evaluation. It is not being maintained here that the best criterion is a doctor's opinion, but rather that no one method can stand out as superior in all cases when dealing with a variety of disorders. The diagnosis is often best established through the integration by an informed individual of data coming from a number of sources. A final problem to be mentioned here is that objective criteria do not yet exist for a number of neurological disorders, but even this problem appears to be undergoing a rapid stage of solution. Most notable in this regard is the *in vivo* differential diagnosis of the degenerative diseases of old age, such as Alzheimer's disease. There is also no objective laboratory marker for multiple sclerosis, and diagnosis of that disorder continues to be made on a clinical basis. Only advances in the neurosciences will lead to ultimate solutions to problems of this type.

In clinical neuropsychology, predictive validity has mainly to do with course of illness. Will the patient get better, stay the same, or deteriorate? Generally, the best way to answer questions of this type is through longitudinal studies, but very few such studies have actually been done. Even in the area of normal aging, in which many longitudinal studies have been accomplished, there have been few

neuropsychologically oriented longitudinal studies. There is, however, some literature on recovery from stroke, much of which is attributable to the work of Meier and collaborators (Meier, 1974). Levin, Benton, and Grossman (1982) provide a discussion of recovery from closed head injury. Of course, it is generally not possible to do a full neuropsychological assessment immediately following closed head injury, and so prognostic instruments used at that time must be relatively simple ones. In this regard, a procedure known as the Glasgow Coma Scale (Teasdale & Jennett, 1974) has well-established predictive validity. Perhaps one of the most extensive efforts directed toward establishment of the predictive validity of neuropsychological tests was accomplished by Paul Satz and various collaborators, involving the prediction of reading achievement in grade school based on neuropsychological assessments accomplished during kindergarten (Fletcher & Satz, 1980; Satz, Taylor, Friel, & Fletcher, 1978). At the other end of the age spectrum, there have been and still are several longitudinal studies contrasting normal elderly individuals with dementia patients (Colsher & Wallace, 1991; Evans et al., 1993). However, we do not yet know from these studies and other ongoing longitudinal investigations what the best prognostic instruments are for predicting the course of dementia or for determining whether or not an elderly individual suspected of having dementia will deteriorate or not. Of great current interest is the "Mild Cognitive Impairment" (MCI) often found in elderly individuals who do not have frank dementia (Aggarwal, Wilson, Beck, Bienias, & Bennett, 2005).

An important aspect of predictive validity has to do with prediction of treatment and rehabilitation outcome. There have been several studies (reviewed by Allen, Goldstein, & Seaton, 1997) concerned with predicting outcome of alcoholism treatment on the basis of neuropsychological test performance. The results of these studies are mixed, but in general it would appear that test performance during the early stages of treatment may bear some relationship to outcome as evaluated by follow-up. Guilmette and Kastner (1996) reviewed prediction of vocational functioning from neuropsychological testing. Before leaving this area, it should be mentioned that there are several not fully documented but apparently reasonable clinical principles related to prediction of treatment outcome. In general, patients with relatively well-circumscribed deficits and perhaps underlying structural lesions, tend to do better in treatment than do patients with more global deficits. There are some data that suggest that early intervention for adults with aphasia, perhaps with 2 months postonset in conjunction with spontaneous recovery is more effective than treatment initiated later (Stein, 1988). Many years ago, Ben-Yishay, Diller, Gertzman, and Gordon (1970) reported that initial level of competence on a task to be trained is related to ability to profit from cues utilized in the training procedure.

In general, studies of predictive validity in neuropsychological assessment have not been as extensive as studies involving concurrent validity. However, the data available suggest that neuropsychological tests can predict degree of recovery or deterioration to some extent and have some capacity to predict treatment outcome. Since many neurological disorders change over time, getting better or worse, and the treatment of neurological disorders is becoming an increasingly active field (Zimmer & Grossberg, 1997), it is often important to have some foreknowledge of

what will happen to the patient in the future in a specific rather than general way and to determine whether or not the patient is a good candidate for some form of treatment. Efforts have also been made to predict functional abilities involved in personal self-care and independent living on the basis of neuropsychological test performance, particularly in the case of elderly individuals (McCue, 1997). The extent to which neuropsychological assessment can provide this prognostic information will surely be associated with the degree of its acceptance in clinical settings.

Studies of the construct validity of neuropsychological tests represent a great amount of the corpus of basic clinical neuropsychological research. Neuropsychology abounds with constructs: short-term memory, attention, visual–spatial skills, psychomotor speed, motor engrams, and cell-assemblies. Tests are commonly characterized by the construct they purport to measure: Test A is a test of long-term memory; Test B is a test of attention; Test C is a test of abstraction ability; Test D is a measure of biological intelligence, etc. Sometimes we fail to recognize constructs as such because they are so well established, but concepts like memory, intelligence, and attention are in fact theoretical entities used to describe certain classes of observable behaviors. Within neuropsychology, the process of construct validation generally begins with an attempt to find a measure that evaluates some concept. Let us begin with a simple example, say the desire to develop a test for memory. Memory, as a neuropsychological construct, would involve a brain–behavior relationship. That is, neuropsychologists are concerned with how impaired brain function affects memory. There are memory tests available, notably the Wechsler Memory Scale fourth edition (WMS-IV) (Wechsler, 2009), but without experimental studies, that scale would only have face validity; that is, it appears to be a test of memory on the basis of the nature of the test items. However, if we ask the related questions, “Does the patient who does well on the scale have a normal memory?”, we would have to know more about the test in regard to how well it assesses memory as a construct. Reasonable alternative hypotheses might be that the scale measures intelligence, educational level or attention, or that these influences confound the test such that impairment of memory per se cannot be unequivocally identified.

The problem may be approached in numerous ways. A factor-analytic strategy may be used in which subtests of the Wechsler Memory Scale are placed into a factor analysis along with educational level and tests of intelligence and attention. It may be found that memory-scale subtests load on their own factor or on factors that receive high loadings from intelligence and attention tests or from educational level. Another approach may involve giving the test to patients with amnesia and to non-amnesic brain-damaged patients. A more sophisticated study may involve administering the Wechsler Memory Scale to these subjects along with other tests. Studies of these types may reveal some of the following hypothetical findings. The Wechsler Memory Scale is highly correlated with IQ, and so it is not possible to tell whether it measures the construct memory specifically or intellectual ability. Some patients cannot repeat stories read to them because they are aphasic and cannot produce words, not because of poor memories. Therefore, interpretation of the measure as an indicator of memory ability cannot be made unequivocally in certain

populations. Certain amnesic patients do exceedingly poorly on certain components of the Wechsler Memory Scale, but well on other components. Such a finding would suggest that memory, as a neuropsychological construct, requires further refinement, since there appears to be a dissociation in patients known to have profound loss of memory between certain memory skills that are intact and others that are severely impaired. Still another approach, suggested by Cronbach (1960), is correlation with practical criteria. Individuals given the Wechsler Memory Scale could be asked to perform a number of tasks, all of which involve practical memory in some way, and the obtained data could be analyzed in terms of what parts of the scale predict success or failure at the various tasks.

Another important way of establishing the construct validity of neuropsychological test batteries involves determining capacity to classify cases into meaningful subtypes. Several groups of investigators have utilized classification statistics, notably R-type factor analysis and cluster analysis in order to determine whether combinations of test scores from particular batteries classify cases in accordance with established diagnostic categories or into types that are meaningful from the standpoint of neuropsychological considerations. A great deal of effort has gone into establishing meaningful, empirically derived subtypes of learning disability (Rourke, 1985), and there has also been work done in the neuropsychologically based empirical classification of neuropsychiatric patients (Goldstein, 1994; Schear, 1987).

It is particularly important to note that the construct validation of neuropsychological tests has involved a multidisciplinary effort with colleagues in cognitive psychology, the experimental psychology of memory and learning (utilizing both human studies and animal models), linguistics, and sensory and perceptual processes. For example, aphasia testing and other forms of language assessment have been profoundly influenced by research in neurolinguistics (Blumstein, 1981; Craiy, Voeller, & Haak, 1988), while memory testing has been correspondingly influenced by a long history of research in information theory and the experimental psychology of memory and learning (Baddeley et al., 1995; Butters & Cermak, 1980). These experimental foundations have aided significantly in the interpretation of clinical tests, and indeed, many new clinical tests are actually derived from laboratory procedures.

## ***Reliability***

While neuropsychological tests should ideally have reliability levels commensurate with other areas of psychometrics, there are some relatively unique problems. These problems are particularly acute when the test-retest method is used to determine the reliability coefficients. The basic problem is that this method really assumes the stability of the subject over testing occasions. When reliability coefficients are established through the retesting of adults over a relatively brief time period, that assumption is a reasonable one, but it is not as reasonable in samples of brain-damaged patients who may be rapidly deteriorating or recovering. Indeed, it is generally thought to be an asset when a test reflects the appropriate changes.

Another difficulty with the test-retest method is that many neuropsychological tests are not really repeatable because of substantial practice effects. The split-half method is seldom applicable, since most neuropsychological tests do not consist of lengthy lists of items, readily allowing for odd-even or other split-half comparisons. In the light of these difficulties, the admittedly small number of reliability studies done with the standard neuropsychological test batteries have yielded perhaps surprisingly good results. [Boll \(1981\)](#) has reviewed reliability studies done with the HRB, and [Goldstein and Watson \(1989\)](#) did a test-retest reliability study with several clinical groups. The test manual ([Golden, Hammeke, & Purisch, 1980](#)) reports reliability data for the Luria-Nebraska Battery. The details of these matters will be discussed later in our reviews of these two procedures. In any event, it seems safe to say that most neuropsychological test developers have not been greatly preoccupied with the reliabilities of their procedures, but those who have studied the matter appear to have provided sufficient data to permit the conclusion that the standard, commonly used procedures are at least not so unreliable as to impair the validities of those procedures.

## An introduction to the comprehensive batteries

There are several generally available comprehensive standard neuropsychological test batteries for adults. The Handbook of Clinical Neuropsychology, published in 1981 ([Filskov & Boll, 1981](#)) only contains chapters on two batteries; the Halstead-Reitan and Luria-Nebraska. [Lezak \(1995\)](#) lists the Halstead-Reitan, the Smith Neuropsychological Battery, and two versions of batteries derived from Luria's work; one by [Christensen \(1975a, 1975b, 1975c\)](#) and [Golden et al.'s \(1980\)](#) Luria-Nebraska (originally South Dakota) Battery. [Jones and Butters \(1983\)](#) reviewed the Halstead-Reitan, Luria-Nebraska, and Michigan batteries. [Benion, Sivan, Hamsher, Varney, and Spreen \(1994\)](#) have produced a manual containing descriptions and instructions for tests these neuropsychologists have been associated with over the years, and [Spreen and Strauss \(1988\)](#) have published a "compendium" of neuropsychological tests, but there was clearly no intention in either case to present these collections of tests as standard batteries. In this chapter, we will consider the Halstead-Reitan and Luria-Nebraska procedures. The Michigan Battery ([Smith, 1975](#)) will not be reviewed, primarily because it consists largely of a series of standardized tests, all of which have their own validity and reliability literature. This literature is thoroughly reviewed by [Lezak \(1995\)](#). We will also consider two new batteries published since the appearance of the previous edition of this book; the Neuropsychological Assessment Battery (NAB; [White & Stern, 2003](#)) and the Meyers Neuropsychological System (2008).

The NAB is a battery containing attention, language, spatial, memory, executive function and screening modules, each of which may be administered independently. It is a standardized, published battery with a manual and materials kit. The Meyers procedure was designed as a flexible battery that assesses the constructs of language

function, fine motor skill, working memory, processing speed, verbal and visual memory, and verbal and visual abstraction and problem solving. The battery contains several tests from the Halstead–Reitan Battery but is intended to be used in a flexible manner. It is quantitatively oriented using many tests incorporated into a database.

The Meyers procedure uses a large number of commonly used tests such as a short form of the WAIS, the Rey Complex Figure Test, the Category Test and the Boston Naming Test. It is actually a scoring and interpretive procedure that uses statistical models (the Rohling interpretive method and profile matching) to produce interpretive material including graphs, profile comparisons and matches between individual patient data and various comparison groups. It overlaps with the HRB in its use of the Wechsler scales, finger tapping and localization, the Trail Making Test and the Category Test, but is not dependent upon giving any particular test or set of tests.

## ***The Halstead–Reitan Battery (HRB)***

### ***History***

The history of this procedure and its founders has been reviewed by Reed (1983) and more recently by Reed and Reed (1997) and by Goldstein and colleagues in this Handbook. These authors trace the beginnings of the battery to the special laboratory established by Halstead in 1935 for the study of neurosurgical patients. The first major report on the findings of this laboratory appeared in a book called Brain and Intelligence: A Quantitative Study of the Frontal Lobes (Halstead, 1947), the title of which suggests that the original intent of Halstead's tests was describing frontal lobe function. In this book, Halstead proposed his theory of "Biological Intelligence" and presented what was, probably, the first factor analysis done with neuropsychological test data. Perhaps more significantly, however, the book contains descriptions of many of the tests now contained in the HRB. As Reed (1983) suggests, the theory of biological intelligence never was widely accepted among neuropsychologists, and the factor analysis had its mathematical problems. But several of the tests that went into that analysis survived, and many of them are commonly used at present. In historical perspective, Halstead's major contributions to neuropsychological assessment, in addition to his very useful tests, include the concept of the neuropsychology laboratory in which objective tests are administered in standard fashions and quantitatively scored, and the concept of the impairment index, a global rating of severity of impairment and probability of the presence of structural brain damage.

Ralph M. Reitan was a student of Halstead at Chicago and was strongly influenced by Halstead's theories and methods. Reitan adopted the methods in the form of the various test procedures and with them established a laboratory at the University of Indiana. He supplemented these tests with a number of additional procedures in order to obtain greater comprehensiveness and initiated a clinical research program that is ongoing, having been taken over by several of his

colleagues after his passing. The program began with a cross-validation of the battery and went on into numerous areas, including validation of new tests added to the battery (e.g., the Trail Making Test), lateralization and localization of function, aging, and neuropsychological aspects of a wide variety of disorders including alcoholism, hypertension, disorders of children, and mental retardation. Theoretical matters were also considered. Some of the major contributions included the concept of type–locus interaction (Reitan, 1966), the analysis of quantitative as opposed to qualitative deficits associated with brain dysfunction (Reitan, 1958, 1959), the concept of the brain-age quotient (Reitan, 1973), and the scheme for levels and types of inference in interpretation of neuropsychological test data (Reitan & Wolfson, 1993). In addition to the published research, Reitan and his collaborators developed a highly sophisticated method of blind clinical interpretation of the HRB that continues to be taught at workshops conducted by students and associates of Dr. Reitan. The HRB, as the procedure came to be known over the years, also has a history.

The HRB has been described as a “fixed battery”, but that is not actually the case. Lezak (1976) says in reference to this development, “This set of tests has grown by accretion and revision and continues to be revised”(p. 440). Halstead’s original battery, upon which the factor analyses were based, included the Carl Hollow Square test, the Dynamic Visual Field Test, the Henmon–Nelson Tests of mental ability, a flicker fusion procedure, and the Time Sense Test. None of these procedures are now widely used, although the Time Sense and Flicker Fusion Tests were originally included in the battery used by Reitan. The tests that survived included the Category Test, the Tactual Performance Test, the Speech Perception Test, and Finger Tapping. Halstead also used the Seashore Rhythm Test, which is included in the current version of the battery, but was not included in the subbattery used by Halstead in his factor analyses. There have been numerous additions, including the various Wechsler Intelligence Scales, the Trail Making Test, a subbattery of perceptual tests, the Reitan Aphasia Screening Test, the Kloev Grooved Pegboard, and other tests that are used in some laboratories but not in others. Alternative methods have also been developed for computing impairment indexes (Reitan, 1991; Russell et al., 1970).

Bringing this brief history into the present, the HRB continues to be widely used, in its entirety or in part, as a clinical and research procedure. Numerous investigators utilize it in their research, and there have been several successful cross-validations done in settings other than Reitan’s laboratory (Goldstein & Shelly, 1972; Vega & Parsons, 1967). In addition to the continuation of factor-analytic work with the battery, several investigators have applied other forms of multivariate analysis to it in various research applications. Several investigations have been conducted relative to objectifying and even computerizing interpretation of the battery, the most well-known efforts probably being the Selz–Reitan rules for classification of brain function in older children (Selz & Reitan, 1979) and the Russell et al. (1970) “Neuropsychological Keys”. The issue of reliability of the battery has been addressed, with reasonably successful results. Clinical interpretation of the battery continues to be taught at workshops and in numerous programs engaged in training

of professional psychologists. The most detailed description of the battery available will be found in [Reitan and Wolfson \(1993\)](#).

Since the publication of the earlier editions of this chapter, much work has been done on the psychometrics of the HRB. We now have available a manual that provides normative information for adults including corrections for age, education, and gender ([Heaton, Grant, & Matthews, 1991](#)), and a number of elegant scoring and rating systems. These include the Neuropsychological Deficit Scale developed by Reitan ([1987, 1991](#)), the Halstead–Russell Neuropsychological Evaluation System (HRNES) ([Russell, 1993](#)), and the Comprehensive Norms for an Extended Halstead–Reitan Battery (CNEHRB) system presented in the Heaton, Grant, and Matthews manual. These systems are all devoted to scaling of the HRB and providing new summary and index scores that are clinically useful. It may be noted that the CNEHRB system has stimulated some controversy revolving around the matter of whether it is appropriate to correct neuropsychological test scores for age and education ([Reitan & Wolfson, 1995](#)).

### **Structure and content**

Although there are several versions of the HRB, the differences tend to be minor, and there appears to be a core set of procedures that essentially all versions of the battery contain. The battery must be administered in a laboratory containing a number of items of equipment and generally cannot be completely administrated at bedside. Various estimates of length of administration are given, but it is probably best to plan on about six to eight hours of patient time. Each test of the battery is independent and may be administered separately from the other tests. However, it is generally assumed that a certain number of the tests must be administered in order to compute an impairment index.

Scoring for the HRB varies with the particular test, such that individual scores may be expressed in time to completion, errors, number correct, or some form of derived score. For research purposes, these scores are generally converted to standard scores so that they may be profiled. [Matthews \(1981\)](#) routinely used a T-score profile in clinical practice, while [Russell et al. \(1970\)](#) rate all of the tests contributing to the impairment index on a six-point scale, the data being displayed as a profile of the ratings. In their system the impairment index may be computed by calculating the proportion of tests performed in the brain-damaged range according to published cutoff scores ([Reitan, 1955](#)) or by calculating the average of the ratings. This latter procedure provides a value called the Average Impairment Rating. [Russell et al. \(1970\)](#) have also provided quantitative scoring systems for the Reitan Aphasia Screening test and for the drawing of a Greek cross that is part of that test. However, some clinicians do not quantify those procedures, except in the form of counting the number of aphasic symptoms elicited. As indicated above, there is the development of a number of additional indices and scoring systems. The General Neuropsychological Deficit Scale (GNDS) ([Reitan & Wolfson, 1993](#)) provides a substantial extension of the Impairment Index and Average Impairment Rating. The system utilizes 42 variables, and is based on four methods of inference; level of

performance, pathognomonic signs, pattern, and right–left differences. It provides both a global score and scores for each inference method category. We will return to other aspects of the battery's structure after the following description of the component tests.

**A. Halstead's biological intelligence tests**

1. *The Halstead Category Test:* This test is a concept-identification procedure in which the subject must discover the concept or principle that governs various series of geometric forms, verbal and numerical material. The apparatus for the test includes a display screen with four horizontally arranged numbered switches placed beneath it. In the original version of the test the stimuli are on slides, and the examiner uses a control console to administer the procedure. The subject is asked to press the switch that the picture reminds him or her of, and is provided with additional instructions to the effect that the point of the test is to see how well he or she can learn the concept idea or principle that connects the pictures. If the correct switch is pressed, the subject will hear a pleasant chime, while wrong answers are associated with a rasping buzzer. The conventionally used score is the total number of errors for the seven groups of stimuli that form the test. Booklet (Adams & Trenton, 1981; DeFillippis, McCampbell, & Rogers, 1979) and abbreviated (Calsyn, O'Leary, & Chaney, 1980; Russell & Levy, 1987; Sherrill, 1987) forms of this test have been developed. The Category Test is currently administered by computer rather than the original apparatus which is no longer available.
2. *The Halstead Tactual Performance Test:* This procedure utilizes a version of the Sequin–Goddard Formboard, but it is done blindfolded. The subject's task is to place all of the 10 blocks into the board, using only the sense of touch. The task is repeated three times, once with the preferred hand, once with the nonpreferred hand and once with both hands, following which the board is removed. After removing the blindfold, the subject is asked to draw a picture of the board, filling in all of the blocks he remembered in their proper locations on the board. Scores from this test include time to complete the task for each of the three trials, total time, number of blocks correctly drawn and number of blocks correctly drawn in their proper locations on the board.
3. *The Speech Perception Test:* The subject is asked to listen to a series of 60 sounds, each of which consists of a double digraph with varying prefixes and suffixes (e.g., geend). The test is given in a four-alternative multiple-choice format, the task being that of underlining on an answer sheet the sound heard. The score is number of errors.
4. *The Seashore Rhythm Test:* This test consists of 30 pairs of rhythmic patterns. The task is to judge whether the two members of each pair are the same as or different from each other and to record the response by writing an S or a D on an answer sheet. The score is either number correct or number of errors.
5. *Finger Tapping:* The subject is asked to tap his or her extended index finger on a typewriter key attached to a mechanical counter. Several series of 10-s trials are run, with both the right and left hand. The scores are average number of taps generally over five trials, for the right and left hand.

**B. Tests added to the battery by Reitan**

1. *The Wechsler Intelligence Scales:* Some clinicians continue to use the Wechsler–Bellevue, some the WAIS, and some the WAIS-R or the latest edition available. In any event, the test is given according to manual instructions and is not modified in any way. It is debatable whether or not HRB users should switch over to

the most recently published editions, currently the WAIS IV, given the substantial differences from earlier versions.

2. *The Trail Making Test:* In Part A of this procedure the subject must connect in order a series of circled numbers randomly scattered over a sheet of 8 1/2 × 11 paper. In Part B, there are circled numbers and letters, and the subject's task involves alternating between numbers and letters in serial order (e.g., 1 to A to 2 to B, etc.). The score is time to completion expressed in seconds for each part.
3. *The Reitan Aphasia Screening Test:* This test serves two purposes in that it contains both copying and language-related tasks. As an aphasia-screening procedure, it provides a brief survey of the major language functions: naming, repetition, spelling, reading, writing, calculation, narrative speech, and right-left orientation. The copying tasks involve having the subject copy a square, Greek cross, triangle, and key. The first three items must each be drawn in one continuous line. The language section may be scored by listing the number of aphasic symptoms or by using the Russell and colleagues' quantitative system. The drawings are not formally scored or are rated through a matching to model system also provided by [Russell et al. \(1970\)](#).
4. *Perceptual Disorders:* These procedures actually constitute a subbattery and include tests of the subject's ability to recognize shapes and identify numbers written on the fingertips, as well as tests of finger discrimination and visual, auditory, and tactile neglect. Number of errors is the score for all of these procedures.

C. Tests Added to the Battery by Others

1. *The Klove Grooved Pegboard Test:* The subject must place pegs shaped like keys into a board containing recesses that are oriented in randomly varying directions. The test is administered twice, once with the right and once with the left hand. Scores are time to completion in seconds for each hand and errors for each hand defined as number of pegs dropped during performance of the task.
2. *The Klove Roughness Discrimination Test:* The subject must order four blocks covered with varying grades of sandpaper presented behind a blind with regard to degree of roughness. Time and error scores are recorded for each hand.
3. *Visual field examination:* [Russell et al. \(1970\)](#) include a formal visual field examination utilizing a perimeter as part of their assessment procedure.

It should be noted that many clinicians, including Reitan and his collaborators, frequently administer a number of additional tests mainly for purposes of assessing personality and level of academic achievement. The MMPI is the major personality assessment method used, and achievement may be assessed with such procedures as the Wide Range Achievement Test-R ([Jastak & Wilkinson, 1984](#)) or the Peabody Individual Achievement Test ([Dunn & Markwardt, 1970](#)). Some clinicians have also adopted the procedure of adding the Wechsler Memory Scale (WMS, WMS-R, WMS-III, WMS IV) to the battery, either in its original form ([Wechsler, 1945, 1987a, 1987b, 1997a, 1997b, 2009](#)) or the Russell modification ([Russell, 1975a](#)). Some form of lateral dominance examination is also generally administered, including tests for handedness, footedness, and eyedness.

D. The "Expanded Halstead-Reitan Battery"

The [Heaton et al. manual \(1991\)](#) contains a number of additional tests that are not a part of the original HRB. They include the Wisconsin Card Sorting Test ([Heaton, 1980](#)), the Story Memory Test (Reitan, unpublished test), the Figure Memory Test (based on the Wechsler Memory Scale; [Wechsler, 1945](#)), the Seashore Tonal Memory Test ([Seashore, Lewis, & Saetewit, 1960](#)), the Digit Vigilance Test ([Lewis & Rennick, 1979](#)), the Peabody Individual Achievement Test ([Dunn & Markwardt, 1970](#)), and the Boston Naming Test ([Kaplan, Goodglass, & Weintraub, 1983](#)).

## Theoretical foundations

There are really two theoretical bases for the HRB, one contained in Brain and Intelligence and related writings of Halstead, the other in numerous papers and chapters written by Reitan and various collaborators (e.g., [Reitan, 1966](#); [Reitan & Wolfson, 1993](#)). They are quite different from each other in many ways, and the difference may be partly accounted for by the fact that Halstead was not primarily a practicing clinician and was not particularly interested in developing his test as psychometric instruments to be used in clinical assessment of patients. Indeed, he never published the tests. He was more interested in utilizing the tests to answer basic scientific questions in the area of brain–behavior relationships in general and frontal lobe function in particular. Reitan's program, on the other hand, can be conceptualized as an effort to demonstrate the usefulness and accuracy of Halstead's tests and related procedures in clinical assessment of brain-damaged patients. It is probably fair to say that Halstead's theory of Biological Intelligence and its factor-analytically based four components (the central integrative field, abstraction, power, and the directional factor), as well as his empirical findings concerning human frontal lobe function have not become major forces in modern clinical neuropsychology. However, they have had, in our view, a more subtle influence on the field.

Halstead was really the first to establish a human neuropsychology laboratory in which patients were administered objective tests, some of which were semiautomated, that utilized standard procedures and sets of instructions. His Chicago laboratory may have been the initial stimulus for the now common practice of trained technician administration of neuropsychological tests. Halstead was also the first to utilize sophisticated, multivariate statistics in the analysis of neuropsychological test data. Even though Reitan did not pursue that course to any great extent, other researchers with the HRB have done so (e.g., [Goldstein & Shelly, 1971, 1972](#)). Thus though the specifics of Halstead's theoretical work have not become well-known and widely applied, the concept of a standard neuropsychological battery administered under laboratory conditions and consisting of objective, quantifiable procedures has made a major impact on the field of clinical neuropsychology. The other, perhaps more philosophical contribution of Halstead was what might be described as his Darwinian approach to neuropsychology. He viewed his discriminating tests as measures of adaptive abilities, as skills that assured human survival on the planet. Many neuropsychologists are now greatly concerned with the relevance of their test procedures to adaptation; the capacity to carry on functional activities of daily living and to live independently ([Sbordone & Long, 1996](#)). This general philosophy is somewhat different from the more traditional models emanating from behavioral neurology in which there is a much greater emphasis on the more medical-pathological implications of behavioral test findings.

Reitan, while always sympathetic with Halstead's thinking, never developed a theoretical system in the form of a brain model or a general theory of the Biological Intelligence type. One could say that Reitan's great concern has always been with the empirical validity of test procedures. Such validity can be established through the collection of large amounts of data obtained from patients with

reasonably complete documentation of their medical/neurological conditions. Both presence and absence of brain damage had to be well documented, and if present, findings related to site and type of lesion had to be established. He has described his work informally as one large experiment, necessitating maximal consistency in the procedures used, and to some extent, the methods of analyzing the data. Reitan and his various collaborators represent the group that was primarily responsible for introduction of the standard battery approach to clinical neuropsychology. It is clear from reviewing the Reitan group's work that there is substantial emphasis on performing controlled studies with samples sufficiently large to allow for application of conventional statistical procedures. One also gets the impression of an ongoing program in which initial findings are qualified and refined through subsequent studies.

It would probably be fair to say that the major thrust of Reitan's research and writings has not been espousal of some particular theory of brain function, but rather an extended examination of the inferences that can be made from behavioral indices relative to the condition of the brain. There is a great emphasis on methods of drawing such inferences in the case of the individual patient. Thus, this group's work has always involved empirical research and clinical interpretation with one feeding into the other. In this regard, there has been a formulation of inferential methods used in neuropsychology (Reitan & Wolfson, 1993) that provides a framework for clinical interpretation. Four methods are outlined: level of performance, pattern of performance, specific behavioral deficits (pathognomonic signs) and right-left comparisons. In other words, one examines for whether or not the patient's general level of adaptive function compares with that of normal individuals, whether there is some characteristic performance profile that suggests impairment even though the average score may be within normal limits, whether there are unequivocal individual signs of deficits, and whether there is a marked discrepancy in functioning between the two sides of the body. As indicated above, these methods have been operationalized in the form of the General Neuropsychological Deficit Scale.

Reitan's theoretical framework is basically empirical, objective, and data-oriented. An extensive research program, by now of almost 70 years' duration, has provided the information needed to make increasingly sophisticated inferences from neuropsychological tests. It thereby constitutes to a significant extent the basis for clinical interpretation. The part of the system that remains subjective is the interpretation itself, but in that regard Reitan (1964) has made the following remark: "Additional statistical methods may be appropriate for this problem but, in any case, progress is urgently needed to replace the subjective-decision-making processes in individual interpretation that presently are necessary" (p. 46). This progress must obviously go beyond the question of presence or absence of brain damage.

### ***Standardization research***

The HRB as a whole meets rigorous validity requirements. Following Halstead's initial validation (1947) it was cross validated by Reitan (1955) and in several other laboratories (Russell et al., 1970; Vega & Parsons, 1967). As indicated above, reviews of validity studies with the HRB have been written over the years by

several authors. Validity in this sense, means that all component tests of the battery that contribute to the impairment index discriminate at levels satisfactory for producing usable cutoff scores for distinguishing between brain-damaged and nonbrain-damaged patients. The major exceptions, the Time Sense and Flicker Fusion Tests, have been dropped from the battery by most of its users. In general, the validation criteria for these studies consisted of neurosurgical and other definitive neurological data. It may be mentioned, however, that most of these studies were accomplished before the advent of the CT and MRI can, and it would probably now be possible to do more sophisticated validity studies, perhaps through correlating extent of impairment with quantitative measures of brain damage obtained using these and other neuroimaging procedures. In addition to what was done with Halstead's tests, validity studies were accomplished with tests added to the battery such as the Wechsler scales, the Trail Making Test, and the Reitan Aphasia Screening Tests, with generally satisfactory results (Reitan, 1966).

By virtue of the level of inferences made by clinicians from HRB data, validity studies must obviously go beyond the question of presence or absence of brain damage. The first issue raised related to discriminative validity between patients with left hemisphere and right hemisphere brain damage. Such measures as Finger Tapping, the Tactual Performance Test, the perceptual disorders subbattery, and the Reitan Aphasia Screening Test all were reported as having adequate discriminative validity in this regard. There have been very few studies, however, that go further and provide validity data related to more specific criteria such as localization and type of lesion. It would appear from one impressive study (Reitan, 1964) that valid inferences concerning prediction at this level must be made clinically, and one cannot call upon the standard univariate statistical procedures to make the necessary discriminations. This study provided the major impetus for Russell et al. (1970) neuropsychological key approach, which was in essence an attempt to objectify higher order inferences.

There is one general area in which the discriminative validity of the HRB was not thought in the past to be particularly robust. The battery does not have great capacity to discriminate between brain-damaged patients and patients with functional psychiatric disorders; notably chronic schizophrenia. There is an extensive literature concerning this matter, but it should be said that some of the research contained in this literature has significant methodological flaws, leaving the findings ambiguous. It may also be pointed out that the construction of the HRB did not have the intention of developing a procedure to discriminate between brain-damaged and schizophrenia patients, and the assumption that it should be able to do so is somewhat gratuitous. Furthermore, Heaton and Crowley (1981) find that with the exception of the diagnosis of chronic schizophrenia, the HRB does a reasonably good job of differential diagnosis. They provided the following conclusion:

*The bulk of the evidence suggests that there is little or no relationship between the degree of emotional disturbance and level of performance on neuropsychological tests. However, significant correlations of this type are sometimes found with schizophrenic groups. (p. 492)*

This matter remains controversial and has become exceedingly complex, particularly since the discovery some time ago of cerebral atrophy in a substantial portion of the schizophrenia population and the development of hypotheses concerning left hemisphere dysfunction in schizophrenia patients (Gruzelier, 1991). Since the last edition there has been a wealth of research concerning brain dysfunction in schizophrenia. Several review articles provide extensive material about this research. Some recent examples include studies of cognitive effort (Whearty et al., 2015), episodic memory (Ragland et al., 2015) and neuropsychological impairment in general (Woodward, 2016). An important study involving the HRB was done some time ago by Braff et al. (1991) clearly demonstrating the presence of cognitive deficits in schizophrenia. The point to be made here is that the user of the HRB should exercise caution in interpretation when asked to use the battery in resolving questions related to differential diagnosis between brain damage and schizophrenia. Some writers have advised the addition of some measure of psychiatric disability, such as the MMPI, when doing such assessments (Russell, 1975b).

Even though there have been several studies of the predictive validity of neuropsychological tests with children (Fletcher & Satz, 1980; Lyon & Flynn, 1991) and other studies with adults that did not utilize the full HRB (Meier, 1974), studies involving formal assessment of the predictive validity of the HRB with regard to adaptive function in adults have only begun to appear recently (Marcotte & Grant, 2009). Within neuropsychology, studies of predictive validity have two aspects: predicting everyday academic, vocational, and social functioning and predicting course of illness. With regard to the former matter, Heaton and Pendleton (1981) document the lack of predictive validity studies using extensive batteries of the HRB type. However, they do report one study (Newman, Heaton, & Lehman, 1978) in which the HRB successfully predicted employment status on 6-month follow-up. With regard to prediction of course of illness, there appears to be a good deal of clinical expertise in this regard, but no major formal studies in which the battery's capacity to predict whether the patient will get better, worse or stay the same are evaluated (Harvey, 2014). This matter is of particular significance in such conditions as head injury and stroke, since outcome tends to be quite variable in these conditions. The changes that occur during those stages are often the most significant ones related to prognosis (e.g., length of time unconscious).

In general, there has not been a great deal of emphasis on studies involving the reliability of the HRB, probably because of the nature of the tests themselves, particularly with regard to the practice effect problem, and because of the changing nature of those patients for whom the battery was developed. Those reliability studies that were done produced satisfactory results, particularly with regard to the reliability to the impairment index (Boll, 1981). The Category Test can have its reliability assessed through the split-half method. In a study accomplished by Shaw (1966), a 0.98 reliability coefficient was obtained.

Norms for the HRB are available in numerous places (Heaton et al., 1991; Reitan & Wolfson, 1993; Russell, 1993; Russell et al., 1970; Vega & Parsons, 1967), but since the battery was never published as a single procedure, there is no published manual that one can refer to for definitive information. Schear (1984) has

published a table of age norms for neuropsychiatric patients. Several laboratories have collected local norms. A great deal is known about the influence of age, education, and gender on the various tests in the HRB, and this information has been consolidated in the [Heaton et al. manual \(1991\)](#). It is somewhat unusual for a procedure in as widespread use as the HRB not to have a commercially published manual. However, detailed descriptions of the procedures as well as instructions for administration and scoring are available in several sources including [Reitan and Wolfson \(1993\)](#), [Jarvis and Barth \(1984\)](#), and [Swercinsky \(1978\)](#).

In summary, the validity of the HRB seems well-established by literally hundreds of studies, including several major cross-validations. These studies have implications for the concurrent, predictive, and construct validity of the battery. Reliability has not received nearly as much attention, but it seems apparent that the battery is sufficiently reliable to not compromise its validity. Extensive age, gender, and education norms have become available ([Heaton et al., 1991](#); [Russell, 1993](#); [Russell et al., 1970](#)) but the relevance of such norms to neuropsychological assessment, particularly with regard to age, is a controversial matter ([Reitan & Wolfson, 1995](#)). There is no commercially available manual for the battery, and so the usual kinds of information generally contained in a manual are not available to the test user in a single place. However, the relevant information is available in a number of separate sources.

## *Evaluation*

Over the course of time that neuropsychological assessment was commonly done, the HRB is without doubt the most widely used standard neuropsychological battery, at least in North America and perhaps throughout the world. In recent years there has been a relative reduction in its clinical use as a comprehensive procedure but portions of it are commonly used. Thus, we chose to continue an emphasis on it in the present edition for this reason and also because some of the new batteries that have recently appeared incorporate many HRB tests or tests like them. Aside from the widespread clinical application of all or portions of the HRB it is used in many multidisciplinary research programs as the procedure of choice for neuropsychological assessment. It therefore has taken on something of a definitive status and is viewed by many experts in the field as the slate-of-the-art instrument for comprehensive neuropsychological assessment. Nevertheless, several criticisms of it have emerged over the years, and some of them will be reviewed here. Each major criticism will be itemized and discussed.

*The HRB is too long and redundant.* The implication of this criticism is that pertinent, clinically relevant neuropsychological assessment can be accomplished in substantially less time than the 6–8 h generally required to administer the full HRB. Other batteries are, in fact, substantially briefer than the HRB. Aside from simply giving fewer or briefer tests another means suggested of shortening neuropsychological assessment is through a targeted, individualized approach rather than through routine administration of a complete battery. The difficulty with this latter alternative is that such an approach can generally only be conducted by an

experienced clinician, and one sacrifices the clinician time and expense that can be saved through administration by trained technicians. The response to the criticism concerning length is generally that shortening of the battery correspondingly reduces its comprehensiveness, and one sacrifices examination of areas that may be of crucial significance in individual cases. Indeed, the battery approach was, in part a reaction to the naivete inherent in the use of single tests for "brain damage." The extent to which the clinician reverts to a single-test approach may reflect the extent to which there is a return to the simplistic thinking of the past. In general, the argument is that to adequately cover what must be covered in a standard comprehensive assessment, the length of the procedure is a necessity. From the point of view of patient comfort and fatigue, the battery can be administered in several sessions over a period of days if necessary.

*The tests in the HRB are insufficiently specific, both in regard to the functions they assess and the underlying cerebral correlates of those functions.* Most of the tests in the battery are quite complex, and it is often difficult to isolate the source of impairment within the context of a single test. Even as apparently simple a procedure as the Speech Perception Test requires not only the ability to discriminate sounds, but to read, make the appropriate written response, and attend to the task. Therefore, failure on the test cannot unequivocally point to a specific difficulty with auditory discrimination. Difficulties of this type are even more pronounced in such highly complex procedures as the Category and Tactual Performance Tests. This criticism eventuates in the conclusion that it is difficult to say anything meaningful about the patient's brain or about treatment because one cannot isolate the specific deficit. In Luria's (1973) terminology one cannot isolate the functional system that is involved, no less the link in that system that is impaired. Failure to do so makes it difficult if not impossible to identify the structures in the brain that are involved in the patient 's impairment as well as to formulate a rehabilitation program, since one doesn't really know in sufficiently specific terms what the patient can and cannot do.

This criticism ideally requires a very detailed response, since it implies a substantially different approach to neuropsychological assessment from the one adopted by developers of the HRB. Perhaps the response can be summarized in a few points. The HRB is founded on empirical rather than on content validity. Inferences are drawn on the basis of pertinent research findings and clinical observations rather than on the basis of what the tests appear to be measuring. The fact that one cannot partial out the various factors involved in successful or impaired performance on the Category Test, for example, does not detract from the significant empirical findings related to this test based on studies of various clinical populations. In any event, Reitan, Hom, and Wolfson (1988) have shown that complex abilities, notably abstraction, are dependent upon the functioning of both cerebral hemispheres and not on a localized unilateral system. The use of highly specific items in order to identify a specific system or system link is a procedure that is closely tied to the syndrome approach of behavioral neurology. Developers of the HRB typically do not employ a syndrome approach for several reasons. First, it depends almost exclusively on the pathognomonic signs method of inference to the

neglect of other inferential methods, and second, the grouping together of specific deficits into a syndrome is felt to be more often in the brain of the examiner than of the patient. The lack of empirical validity of the so-called Gerstmann Syndrome is an example of this deficiency in this particular approach (Benton, 1961). Another major point is that the HRB is a series of tests in which interpretation is based not on isolated consideration of each test taken one at a time, but on relationships among performances on all of the tests. Therefore, specific deficits can be isolated, in some cases at least, through interest comparisons rather than through isolated examination of a single test.

Returning to our example, the hypothesis that there is impairment on the Speech Perception Test because of failure to read the items accurately can be evaluated through looking at the results of the aphasia screening or reading-achievement test given. Finally, complex tests are likely to have more ecological validity than simple tests of isolated abilities. Thus, the Category Test or Tactual Performance Test results can tell the clinician more about real-world functioning than can the simpler tests. Simple tests were developed in the context of neurological diagnosis, while the tests in the HRB seem more oriented to assessing adaptive functioning in the environment.

*The HRB is not sufficiently comprehensive, particularly in that it completely neglects the area of memory.* The absence of formal memory testing in this battery has been noted by many observers and appears to be a valid criticism. On the face of it, it would appear that the battery would be incapable of identifying and providing meaningful assessments of patients with pure amnesic syndromes (e.g., patients with Korsakoff's syndrome). The absence of formal memory testing as part of the HRB is something of a puzzlement; although memory is involved in many of the tests, it is difficult to isolate the memory component as a source of impairment. Such isolation is readily achieved through such standard, commonly available procedures as list or paired-associate learning.

We know of no formal response to this criticism, but the point of view could be taken that pure amnesic syndromes are relatively rare, and the HRB would probably not be the assessment method of choice for many of the rarely occurring specific syndromes. We would view this response as weak in view of the recently reported significance of memory defect in a number of disorders (Baddeley et al., 1995). Apparently, Halstead did not work with patients of those types, particularly patients with Alzheimer's and Huntington's disease, and so may have failed to note the significance of memory function in those disorders. However, this criticism is probably the one most easily resolved, since all that is required is addition of some formal memory testing to the battery. Such tests are included in the CNEHRB and the HRNES.

*The HRB cannot discriminate between brain-damaged and schizophrenia patients.* This matter has already been discussed, and most of the evidence (Heaton & Crowley, 1981) indicates that the performance of chronic schizophrenia patients on the HRB may be indistinguishable from that of patients with generalized, structural brain damage. There are essentially two classes of response to this criticism. First, there is a disclaimer that the HRB was never designed for this kind of

differential diagnosis, and so it is not surprising that it fails when it is inappropriately used for that purpose. Second, and perhaps much more significantly, is the finding that many individuals with schizophrenia have clearly identifiable brain abnormalities as assessed by CT and MRI scans, and tests of the HRB type can now be viewed as accurately identifying the behavioral correlates of that condition (Marsh, Lauriello, Sullivan, & Pfefferbaum, 1996). Furthermore, there are now several studies that indicate that schizophrenia is a neuropsychologically heterogeneous condition, and that there is a lack of relationship between neuropsychological test results and psychiatric diagnosis in the case of several psychiatric disorders (Goldstein, 1994; Townes et al., 1985).

*Findings reported from Reitan's laboratory cannot be replicated in other settings.* Here we have particular reference to the criticisms raised by Smith of Reitan's early Wechsler–Bellevue laterality studies. In a series of papers, Smith (1965, 1966a, 1966b) presented empirical and theoretical arguments against the reported finding that patients with left hemisphere lesions have lower verbal than performance IQs on the Wechsler–Bellevue, while the reverse was true for patients with right hemisphere brain damage. Smith was unable to replicate these findings in patients with lateralized brain damage that he had Wechsler–Bellevue data available on and also presented theoretical arguments against the diagnostic and conceptual significance of this finding. Kjove (1974) analyzed the Smith versus Reitan findings in terms of possible age and neurological differences between the studies. Reviewing the research done to the time of writing, he also concluded that most of the research, with Smith as the only pronounced exception, essentially confirmed Reitan's original findings.

*In recent years, it has been asserted informally that the HRB is “old-fashioned” and out of date because it has not kept up with developments in neuroscience, experimental neuropsychology, and psychometrics. So-called process and cognitive approaches now reflect the state-of-the-art in neuropsychological assessment.* This view is not supported for a number of reasons including continued widespread clinical use of all or some of the HRB, great popularity of continuing education concerning clinical use of the HRB, and a large number of publications in the literature written by Reitan shortly before his death and collaborators themselves (e.g., Reitan & Wolfson, 1995, 1997) or by others using HRB-oriented test batteries (e.g., Goldstein, Beers, & Shemansky, 1996; Goldstein & Shemansky, 1997; Palmer et al., 1997).

In summary, many criticisms have been raised of the HRB as a comprehensive, standard neuropsychological assessment system. While pertinent and reasonable responses have been made to most or all of these critiques, members of the profession have nevertheless sensed in recent years the desire to develop alternative procedures. Despite the pertinent replies to criticisms, there appear to be many clinicians who still feel that the HRB is too long, does neglect memory, and in many cases is insufficiently specific. Some holders of these views adopted an individualized approach, or modified the HRB, while others sought alternative standard batteries. Two new comprehensive batteries have appeared, the NAB and the Meyers Neuropsychological System that are now in common use. However, it is noted that these systems are heavily influenced by the HRB.

## **The Luria–Nebraska Neuropsychological Battery**

### **History**

This procedure, previously known as the Luria-South Dakota Neuropsychological Battery or as A Standard Version of Luria's Neuropsychological Tests, was first reported on in 1978 (Golden, Hammek, & Purisch, 1978; Purisch, Golden, & Hammeke, 1978) in the form of two initial validity studies. One could provide a lengthy history of this procedure, going back to Luria's original writings, or a brief one only recording events that occurred since the time of preparation of the two publications cited above. We will take the latter alternative, for reasons that will become apparent. Prior to the past quarter of a century, Luria was a shadowy figure to most English-speaking neuropsychologists. It was known that he was an excellent clinician who had developed his own methods for evaluating patients as well as his own theory, but the specific contents were unknown until translations of some of his major works appeared in the 1960s (e.g., Luria, 1966). However, when these works were read by English-speaking professionals, it became apparent that Luria did not have a standard battery of the HRB type and did not even appear to use standardized tests. Thus, while his formulations and case presentations were stimulating and innovative, nobody quite knew what to do with these materials in terms of practical clinical application. One alternative, of course, was to go to the Soviet Union and study with Luria, and, in fact, Anne-Lise Christensen did just that and reported what she had learned in a book called Luria's Neuropsychological Investigation (Christensen, 1975a). The book was accompanied by a manual and a kit containing test materials used by Luria and his coworkers (Christensen, 1975b, 1975c). Even though some of Luria's procedures previously appeared in English in Higher Cortical Functions (1966) and Traumatic Aphasia (1970), they were never presented in a manner that encouraged direct administration of the test items to patients. Thus, the English-speaking public had in hand a manual and related materials that could be used to administer some of Luria's tests. These materials did not contain information relevant to standardization of these items. There are no scoring systems, norms, data regarding validity and reliability, or review of research accomplished with the procedure as a standard battery. This work was taken on by a group of investigators under the leadership of Charles J. Golden and was initially reported on in the two 1978 papers cited above. Thus, in historical sequence, Luria adopted or developed these items over the course of many years, Christensen published them in English but without standardization data, and finally Golden and collaborators provided quantification and standardization. Since that time, Golden's group as well as other investigators have produced a massive amount of studies with what is now known as the Luria–Nebraska Neuropsychological Battery. The battery was originally published in 1980 by Western Psychological Services (Golden et al., 1980) and is now used in clinical and research applications. An alternate form of the battery is available (Golden et al., 1985), as is a children's version (Golden, 1981).

## ***Structure and content***

The basic structure of the Luria–Nebraska in its current version contains 269 items, each of which may be scored on a two- and three-point scale. A score of 0 indicates normal performances. Some items may receive a score of 1 indicating borderline performance. A score of 2 indicates clearly abnormal performance. The items are organized into the categories provided in the Christensen kit (Christensen, 1975c), but while Christensen organized the items primarily to suggest how they were used by Luria, in the Luria–Nebraska version the organization is presented as a set of quantitative scales. The raw score for each scale is the sum of the 0, 1, and 2 item scores. Thus, the higher the score, the poorer the performance. Since the scales contain varying numbers of items, raw scale scores are converted to T scores with a mean of 50 and a standard deviation of 10. These T scores are displayed as a profile on a form prepared for that purpose. The scores for the individual items may be based on speed, accuracy, or quality of response. In some cases, two scores may be assigned to the same task, one for speed and the other for accuracy. These two scores are counted as individual items. For example, one of the items is a block-counting task, with separate scores assigned for number of errors and time to completion of the task. In the case of time scores, blocks of seconds are associated with the 0, 1, and 2 scores. When quality of response is scored, the manual provides both rules for scoring, and, in the case of copying tasks, illustrations of figures representing 0, 1, and 2 scores.

The 269 items are divided into 11 content scales, each of which is individually administrable. These scales were originally called the Motor, Rhythm, Tactile, Visual, Receptive Speech, Expressive Speech, Writing, Reading, Arithmetic, Memory and Intellectual Processes scales. In the new manual, the names of the content scales have been replaced by abbreviations. Thus, the clinical scales are referred to as the C1–C11 scales. In addition to these 11 content scales, there are three derived scales that appear on the standard profile form: the Pathognomonic, Left Hemisphere, and Right Hemisphere scales. The Pathognomonic scale contains items from throughout the battery found to be particularly sensitive to presence or absence of brain damage.

Several other scales have been developed by Golden and various collaborators, all of which are based on different ways of scoring the same 269 items. These special scales include empirically derived right and left hemisphere scales (McKay & Golden, 1979a), a series of localization scales (McKay & Golden, 1979b), a series of factor scales (McKay & Golden, 1981), and double discrimination scales (Golden, 1979). The empirical right and left hemisphere scales contain items from throughout the battery and are based on actual comparisons among patients with right and left hemisphere, and diffuse brain damage. The localization scales are also empirically derived (McKay & Golden, 1979b), being based on studies of patients with localized brain lesions. There are frontal, sensory-motor, temporal, and parieto-occipital scales for each hemisphere. The factor scales are based on extensive factor-analytic studies of the battery involving factor analyses of each of the major content scales (e.g., Golden & Berg, 1983). The empirical right and left

hemisphere, localization, and factor scales may all be expressed in T scores with a mean of 50. The double discrimination scales which have been shown to be effective in diagnosis of multiple sclerosis (Golden, 1979), involve development of two scales: one contains items on which patients with a particular diagnosis do worse than the general neurological population, and the other contains items on which patients do better. Classification to the specific group is made when scores are in the appropriate range on both scales. There are also two scales that provide global indices of dysfunction, and are meant as equivalents to the Halstead Impairment Index. They are called the Profile Elevation and Impairment scales.

The Luria–Nebraska procedure involves an age and education correction. It is accomplished through computation of a cutoff score for abnormal performance based on an equation that takes into consideration both age and education. The computed score is called the critical level and is equal to  $0.214 \text{ (age)} + 1.47 \text{ (education)} + 68.8 \text{ (constant)}$ . Typically, a horizontal line is drawn across the profile at the computed critical level point. The test user has the option of considering scores above the critical level, which may be higher or lower than 60, as abnormal.

As indicated above, extensive factor-analytic studies have been accomplished, and the factor structure of each of the major scales has been identified. These analyses were based on item intercorrelations, rather than on correlations among the scales. It is important to note that most items on any particular scale correlate more highly with other items on that scale than they do with items on other scales (Golden, 1981). This finding lends credence to the view that the scales are at least somewhat homogeneous, and thus that the organization of the 269 items into those scales can be justified.

### *Theoretical foundations*

As in the case of the HRB, one could present two theoretical bases for the Luria–Nebraska, one revolving around the name of Luria and the other around the Nebraska group, Golden and his collaborators. This view is elaborated upon in Goldstein (1986a, 1986b). It is to be noted in this regard that Luria himself had nothing to do with the development of the Luria–Nebraska Battery, nor did any of his coworkers. The use of his name in the title of the battery is, in fact, somewhat controversial and seems to have been essentially honorific in intent, recognizing his development of the items and the underlying theory for their application. Indeed, Luria died some time before publication of the battery but was involved in the preparation of the Christensen materials, which he endorsed. Furthermore, the method of testing employed by the Luria–Nebraska was not Luria's method, and the research done to establish the validity, reliability, and clinical relevance of the Luria–Nebraska was not the kind of research done by Luria and his collaborators. Therefore, our discussion of the theory underlying the Luria–Nebraska Battery will be based on the assumption that the only connecting link between Luria and that procedure is the set of Christensen items. In doing so, it becomes clear that the basic theory underlying the development of Luria–Nebraska is based on a philosophy of science that stresses empirical validity, quantification, and application of

established psychometric procedures. Indeed, as pointed out elsewhere ([Goldstein, 1986a, 1986b](#)), it is essentially the same epistemology that characterizes the work of the Reitan group.

The general course charted for establishment of quantitative, standard neuropsychological assessment batteries involves several steps: (1) determining whether the battery discriminates between brain-damaged patients in general and normal controls; (2) determining whether it discriminates between patients with structural brain damage and conditions that may be confused with structural brain damage, notably various functional psychiatric disorders; (3) determination of whether the procedure has the capacity to lateralize and regionally localize brain damage; and (4) determination of whether there are performance patterns specific to particular neurological disorders, such as alcoholic dementia or multiple sclerosis. In proceeding along this course, it is highly desirable to accomplish appropriate cross-validations and to determine reliability. This course was taken by Golden and his collaborators, in some cases with remarkable success. Since the relevant research was accomplished during relatively recent years, it had the advantages of being able to benefit from the new brain-imaging technology, notably the CT scan, and the application of high-speed computer technologies, allowing for extensive use of powerful multivariate statistical methods. With regard to methods of clinical inference, the same methods suggested by Reitan level of performance, pattern of performance, pathognomonic signs, and right-left comparisons are the methods generally used with the Luria-Nebraska.

Adhering to our assumption that the Luria-Nebraska bears little resemblance to Luria's methods and theories, there seems little point in examining the theoretical basis for the substance of the Luria-Nebraska Battery. For example, it seems that there would be little point in examining the theory of language that underlies the Receptive Speech and Expressive Speech scales or the theory of memory that provides the basis for the Memory Scale. An attempt to produce such an analysis was made some time ago by [Spiers \(1981\)](#), who examined the content of the Luria-Nebraska scales and evaluated it with reference not so much to Luria's theories, but to current concepts in clinical neuropsychology in general. However, despite the thoroughness of the Spiers review, it seems to miss the essential point that the Luria-Nebraska is a procedure based primarily on studies of empirical validity. One can fault it on the quality of its empirical validity, but not on the basis that it utilizes such an approach.

It therefore appears that the Luria-Nebraska Battery does not constitute a means of using Luria's theory and methods in English-speaking countries. It is a standardized psychometric instrument with established validity using Luria's theory and methods. The choice of using items selected by [Christensen \(1975b\)](#) to illustrate Luria's testing methods was, in retrospect, probably less crucial than the research methods chosen to investigate the capabilities of this item set. Indeed, it is somewhat misleading to characterize these items as "Luria's tests," since many items selected by [Christensen \(1975b\)](#) to illustrate Luria's testing methods were, in retrospect, probably less crucial than the research methods chosen to investigate the capabilities of this item set. Surely, one cannot describe asking a patient to interpret

proverbs or determine two-point thresholds as being exclusively “Luria’s tests.” They are, in fact, venerable, widely used procedures. The Luria–Nebraska is a standardized psychometric instrument with established validity for certain purposes and reliability.

### ***Standardization research***

Fortunately, there are published manuals for the Luria–Nebraska (Golden et al., 1980, 1985) that describe the battery in detail and provide pertinent information relative to validity, reliability, and norms. There are also several review articles (e.g., Golden, 1981; Moses & Purisch, 1997; Purisch & Shordone, 1986) that comprehensively describe the research done with the battery. Very briefly reviewing this material, satisfactory discriminative validity has been reported in studies directed toward differentiating miscellaneous brain-damaged patients from normal controls and from chronic schizophrenia. Cross-validations were generally successful, but Shelly and Goldstein (1983) could not fully replicate the studies involved with discriminating between brain-damaged and schizophrenia patients. Discriminative validity studies involving lateralization and localization achieved satisfactory results, but the localization studies were based on small samples. Quantitative indices from the Luria–Nebraska were found to correlate significantly with CT-scan quantitative indices in alcoholism (Golden et al., 1981) and schizophrenia (Golden et al., 1980) samples. There have been several studies of specific neurological disorders including multiple sclerosis (Golden, 1979), alcoholism (Chmielewski & Golden, 1980), Huntington’s disease (Moses, Golden, Berger, & Wisniewski, 1981) and learning-disabled adults (McCue, Shelly, Goldstein, & Katz-Garris, 1984), all with satisfactory results in terms of discrimination.

The test manual reports reliability data. Test–retest reliabilities for the 13 major scales range from 0.78 to 0.96. The problem of interjudge reliability is generally not a major one for neuropsychological assessment, since most of the tests used were quite objective and have quantitative scoring systems. However, there could be a problem with the Luria–Nebraska, since the assignment of 0, 1, and 2 scores sometimes requires a judgment by the examiner. During the preliminary screening stage in the development of the battery, items in the original pool that did not attain satisfactory interjudge reliability were dropped. A 95% interrater agreement level was reported by the test constructors for the 282 items used in an early version of the battery developed after the dropping of those items. The manual contains means and standard deviations for each item based on samples of control, neurologically impaired, and schizophrenia subjects. An alternate form of the battery is available. There have also been a small number of successful predictive or ecological validity studies reviewed in Moses and Purisch (1997). It is unclear whether or not there have been studies addressed to the issue of construct validity. Stambrook (1983) suggested that studies involved with item-scale consistency, factor analysis, and correlation with other instruments are construct-validity studies, but it does not appear to us that they are directed toward validation of Luria’s constructs. The attempt to apply Luria’s constructs has not in fact involved the empirical testing of

specific hypotheses derived from Luria's theory. Thus, we appear to have diagnostic or discriminative validity established by a large number of studies. There also seems to be content validity, since the items correlate most highly with the scale to which they are assigned, but the degree of construct validity remains unclear. For example, there have been no studies of Luria's important construct of the functional system or of his hypotheses concerning the role of frontal lobe function in the programming, regulation, and verification of activity (Luria, 1973).

With regard to Form II, it is important to note that Moses and Purisch (1997) have provided clear evidence that Forms I and II are not equivalent forms, and should not be used in longitudinal studies as alternate forms. Form II cannot be hand-scored and must be scored using a computer program. It also includes a new clinical scale; Intermediate Memory (C12).

### ***Evaluation***

At the time of the past edition, the early heated controversies concerning the Luria–Nebraska Battery appear to have diminished and we no longer see the highly critical reviews that appeared shortly after the procedure first appeared. At that time Adams (1980) criticized it primarily on methodological grounds, Spiers (1981) on the basis that it was greatly lacking in its capacity to provide a comprehensive neuropsychological assessment, Crosson and Warren (1982) because of its deficiencies with regard to assessment of aphasia and aphasic patients, and Stambrook (1983) on the basis of a number of methodological and theoretical considerations. Replies were written to several of these reviews (e.g., Golden, 1980), and a rather heated literature controversy eventuated. This literature was supplemented by several case studies (e.g., Delis & Kaplan, 1982) in which it was shown that the inferences that would be drawn from the Luria–Nebraska were incorrect with reference to documentation obtained for those cases. These criticisms can be divided into general and specific ones. Basically, there are two general criticisms: (1) The Luria–Nebraska Battery does not reflect Luria's thinking in any sense, and his name should not be used in describing it; and (2) there are several relatively flagrant methodological difficulties involved in the standardization of the procedure. The major specific criticisms primarily involve the language related and memory scales. With regard to aphasia, there are essentially two points. First, there is no system provided, nor do the items provide sufficient data to classify the aphasias in terms of some contemporary system (e.g., Goodglass & Kaplan, 1983). Second, the battery is so language-oriented that patients with aphasia may fail many of the non-language tasks because of failure to comprehend the test instructions or to make the appropriate verbal responses indicative of a correct answer. For example, on the Tactile scale, the patient must name objects placed in the hands. Patients with anomia or anomic aphasia will be unable to do that even though their tactile recognition skills may be perfectly normal. With regard to memory, the Memory Scale is criticized because of its failure to provide a state-of-the-art comprehensive memory assessment (Russell, 1981). Golden has responded to this criticism through adding additional items involving delayed recall to the alternate form of the battery. Moses

and Purisch (1997) have reviewed this critical material and concluded that for various reasons several of the criticisms were unfounded, the validity and reliability of the Luria–Nebraska has been demonstrated in a large number of studies, and that efforts were made in updated versions of the battery to correct for reasonable criticisms.

In providing an evaluation of the Luria–Nebraska, one can only voice an opinion, as others have, since its existence has stimulated a polarization into “those for it” and “those against it.” We would concur with Stambrook’s view (1983), which essentially is that it is premature to make an evaluation and that major research programs must be accomplished before an informed opinion can be reached. We continue to hold this opinion at the present writing, many years after appearance of the Stambrook paper. However, the need for an expanded database expressed in the previous versions of this chapter has been largely fulfilled through the efforts of James Moses and collaborators (Moses & Purisch, 1997). There is still need for more evaluation of the actual constructs on which the procedure is based and assessment of its clinical usefulness relative to other established procedures such as the HRB or individual approaches. The following remark by Stambrook (1983) continues to reflect a highly reasoned approach to this issue. “The clinical utility of the LNNB does not depend upon either the publisher’s and test developer’s claims, or on conceptual and methodological critiques, but upon carefully planned and well-executed research” (p. 266). Various opinions have also been raised with regard to whether it is proper to utilize the Luria–Nebraska in clinical situations. It continues to be my view of the matter that it may be so used as long as inferences made from it do not go beyond what can be based on the available research literature. In particular, the test consumer should not be led to believe that administration and interpretation of the Luria–Nebraska battery provide an assessment of the type that would have been conducted by Luria and his coworkers, or that one is providing an application of Luria’s method. The procedure is specifically not Luria’s method at all, and the view that it provides valid measures of Luria’s constructs and theories has not been verified. Even going beyond that point, attempts to verify some of Luria’s hypotheses (e.g., Drewe, 1975; Goldberg & Tucker, 1979) have not always been completely successful. Therefore, clinical interpretations, even when they are based on Luria’s actual method of investigation, may be inaccurate because of inaccuracies in the underlying theory.

### *Other fixed battery approaches*

Often times, relatively rapid screening is needed to assess cognitive function that is more extensive than mental status type assessments. Several batteries have been developed for this purpose, which are often truncated portions of full fledged batteries. One such popular battery is the Repeatable Battery for Assessment of Neuropsychological Status (RBANS) (Randolph, 1998). The RBANS was constructed from tests familiar to most neuropsychologist. It contains ten subtests which result in five index scores: Language, Attention, Visuospatial/constructional, Immediate Memory and Delayed Memory, as well as a summary measure. Each

index results in a standard score with a mean of 100 + 15. The RBANS is often used for testing that requires a limited amount of time, such as inpatient evaluations where comprehensive assessments may not be practical.

More recently, short fixed batteries have been developed that are specific to particular disease entities. Examples include the Minimal Assessment of Cognitive Function (MACFIMS) for cognitive assessment in multiple sclerosis (Benedict et al., 2006) and the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) battery for assessment in schizophrenia (Green et al., 2004). Both of these batteries were constructed by a team of experts by consensus to assess the most important domains of cognitive function in these disease entities. Existing validated and standardized tests for each of the domains identified were selected to be included in the battery. For example, for MATRICS, tests selected for the domains of cognitive deficit in schizophrenia are working memory, attention/vigilance, verbal learning and memory, visual learning and memory, reasoning and problem solving, speed of processing, and social cognition. For MACFIMS, existing tests were selected to assess processing speed, verbal and visual learning and memory, executive functions and visual perception. They are described as “comprehensive” in that they assess the primary areas of cognition deemed to be most important for the particular disease entity in question. They are not designed to be utilized outside its initially intended purpose.

### ***Flexible batteries***

The previous sections outlined what has been termed “fixed batteries” which consist of a set number of tests administered to all patients. In contrast, the flexible battery approach allows the clinician “flexibility” to choose tests that directly address a referral question. Some flexible batteries, such as the Meyers Neuropsychological Battery (Meyers & Rohling, 2004) begin with a core set of tests from which other tests are added. The additional tests are added by the clinician based on the clinical need for elaboration regarding the specific assessment being conducted. While there is flexibility in test selection, administration is conducted in accordance within standardized procedures. Strength of this approach is its flexibility to address specific questions, and the allowance for hypothesis testing by the clinician. Challenges to this approach include a dependence on the clinician’s clinical experience and the lack of standardization which may challenge test–retest requirements, such as when follow-up testing is conducted by another neuropsychologist.

It should be recognized that a flexible approach to neuropsychological assessment is not new as early pioneers in neuropsychological assessment utilized and advocated for such techniques (e.g., Benton, 1977). Also, the flexible battery should not be confused with the flexible or process approach. In this approach, a clinician may administer just parts of a fixed battery or even “test the limits” by altering standardized administration (e.g., allowing additional time) to examine particular aspects of behavior. The Boston Process Approach to Neuropsychological assessment (Ashendorf, Swenson & Libon, 2013) is an example of the flexible approach to testing.

## Conclusions

In the first part of this chapter, general problems in the area of standardization of comprehensive neuropsychological test batteries were discussed, while the second part contained brief reviews of the two most widely used procedures, the HRB and the Luria–Nebraska. There is increasing use of the NAB and the Meyers Neuropsychological System.

It was generally concluded that the HRB and Luria–Nebraska batteries have their advantages and disadvantages. The HRB is well established and detailed but is lengthy, cumbersome, and neglects certain areas, notably memory. The Luria–Nebraska is also fairly comprehensive and briefer than the HRB but is currently quite controversial and is thought to have major deficiencies in standardization and rationale, at least by some observers. We have taken the view that all of these standard batteries are screening instruments, but not in the sense of screening for presence or absence of brain damage. Rather, they may be productively used to assess a number of functional areas such as memory, language, or visual–spatial skills that may be affected by brain damage. With the development of the new imaging techniques in particular, it is important that the neuropsychologist not simply tell the referring agent what he or she already knows. The unique contribution of standard neuropsychological assessment is the ability to describe functioning in many crucial areas on a quantitative basis. The extent to which one procedure can perform this type of task more accurately and efficiently than other procedures will no doubt greatly influence the relative acceptability of these batteries by the professional community.

The NAB and Meyers Neuropsychological System constituting the two new comprehensive procedures that appeared after the publication of this book's previous editions use the same or similar tests used over a long period of time in neuropsychological assessment, but provide organizational systems and standardization information that promotes analysis and interpretation. They are therefore both derived from previously developed assessment methods but provide systematic methods of administering, scoring and interpreting results of these procedures. They do not yet have the history of neuropsychological research that forms the basis for the HRB and, to some extent, the Luria–Nebraska. They are presented as flexible systems because not all of the tests contained in these procedures are typically administered to individual patients, but the clinician has the option of routinely using the same tests with all patients, creating what might be characterized as an individualized standard battery. Both of these new procedures are strongly psychometrically oriented with strong consideration given to normative samples, validity and reliability. The influence of the HRB and related procedures is noted in both of the new methods of assessment.

## References

- Adams, K. M. (1980). In search of Luria's battery: A false start. *Journal of Consulting and Clinical Psychology*, 48, 511–516.

- Adams, R. L., & Trenton, S. L. (1981). Development of a paper-and-pen form of the Halstead category test. *Journal of Consulting and Clinical Psychology*, 49, 298–299.
- Aggarwal, N. T., Wilson, R. S., Beck, T. L., Bienias, J. L., & Bennett, D. A. (2005). Mild cognitive impairment in differential functional domains and incident Alzheimer's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 76, 1479–1484.
- Albert, M. L., Goodglass, H., Helm, N. A., Ruben, A. B., & Alexander, M. P. (1981). *Clinical aspects of dysphasia*. New York: Springer-Verlag/Wein.
- Allen, D. N., Goldstein, G., & Seaton, B. E. (1997). Cognitive rehabilitation of chronic alcohol abusers. *Neuropsychology Review*, 7, 21–39.
- Ashendorf, L., Swenson, R., & Libon, D. J. (2013). *The Boston Process Approach to Neuropsychological Assessment: A practitioner's Guide*. New York: Oxford University Press.
- Baddeley, A. D., Wilson, B. A., & Watts, F. N. (1995). *The handbook of memory disorders*. Chichester, UK: Wiley.
- Bender, L. (1938). A visual–motor Gestalt test and its clinical use. *American Psychiatric Association, Research Monographs*, No. 3.
- Bender, M. B. (1952). *Disorders in perception*. Springfield, IL: Charles C. Thomas.
- Benedict, R. H., Cookfair, D., Gavett, R., Gunther, M., Munschauer, F., Garg, N., & Weinstock-Guttman, B. (2006). Validity of the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society*, 12(4), 549–558.
- Benson, D. F., & Ardila, A. (1996). *Aphasia: A clinical perspective*. New York: Oxford University Press.
- Benton, A. L. (1961). The fiction of the Gerstmann syndrome. *Journal of Neurology, Neurosurgery and Psychiatry*, 24, 176–181.
- Benton, A. L. (1963). *The revised visual retention test*. New York: Psychological Corporation.
- Benton, A. L. (1977). Psychological testing. In A. B. Baker, & L. H. Baker (Eds.), *Clinical neurology*. New York: Harper & Row.
- Benion, A. L., Sivan, A. B., Hamsher, K. D., Varney, N. R., & Spreen, O. (1994). *Contributions to neuropsychological assessment* (2nd ed.). New York: Oxford University Press.
- Ben-Yishay, Y., Diller, L., Gertsman, L., & Gordon, W. (1970). Relationship between initial competence and ability to profit from cues in brain-damaged individuals. *Journal of Abnormal Psychology*, 78, 248–259.
- Blumstein, S. E. (1981). Neurolinguistic disorders: Language–brain relationship. In S. B. Filskov, & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (pp. 227–256). New York: Wiley-Interscience.
- Boll, T. J. (1981). The Halstead–Reitan neuropsychology battery. In S. B. Filskov, & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (pp. 577–607). New York: Wiley-Interscience.
- Braff, D. L., Heaton, R., Kuck, J., Cullum, L., Moranville, J., Grant, I., ... Zisook, S. (1991). The generalized pattern of neuropsychological deficits in outpatients with chronic schizophrenia with heterogeneous Wisconsin Card Sorting Test results. *Archives of General Psychiatry*, 48, 891–898.
- Butters, N. (August, 1983). *Clinical aspects of memory disorders: Contributions from experimental studies of amnesia and dementia*. Presented at American Psycholocial Association, Anaheim, CA.
- Butters, N. M., & Cermak, L. S. (1980). *Alcoholic Korsakoff's syndrome*. New York: Academic Press.

- Calsyn, D. A., O'Leary, M. R., & Chaney, E. F. (1980). Shortening the category test. *Journal of Consulting and Clinical Psychology*, 48, 788–789.
- Canter, A. (1970). *The Canter background interference procedure for the Bender–Gestalt test: Manual for administration, scoring and interpretation*. Iowa City, IA: Iowa Psychopathic Hospital.
- Chmielewski, C., & Golden, C. J. (1980). Alcoholism and brain damage: An investigation using the Luria–Nebraska neuropsychological battery. *International Journal of Neuroscience*, 10, 99–105.
- Christensen, A. L. (1975a). *Luria's neuropsychological investigation*. New York: Spectrum.
- Christensen, A. L. (1975b). *Luria's neuropsychological investigation: Manual*. New York: Spectrum.
- Christensen, A. L. (1975c). *Luria's neuropsychological investigation: Test cards*. New York: Spectrum.
- Colsher, P. L., & Wallace, R. B. (1991). Longitudinal application of cognitive function measures in a defined population of community-dwelling elders. *Annals of Epidemiology*, 1, 215–230.
- Craig, M. A., Voeller, K. K. S., & Haak, N. J. (1988). Questions of developmental neurolinguistic assessment. In M. G. Tramontana, & S. R. Hooper (Eds.), *Assessment issues in child neuropsychology* (pp. 249–279). New York: Plenum Press.
- Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper & Brothers.
- Crosson, B., & Warren, R. L. (1982). Use of the Luria–Nebraska neuropsychological battery in aphasia: A conceptual critique. *Journal of Consulting and Clinical Psychology*, 50, 22–31.
- Cummings, J. L. (Ed.), (1990). *Subcortical dementia*. New York: Oxford University Press.
- DeFillippis, N. A., McCampbell, E., & Rogers, P. (1979). Development of a booklet form of the category test: Normative and validity data. *Journal of Clinical Psychology*, 50, 32–39.
- Delis, D. C., & Kaplan, E. (1982). The assessment of aphasia with the Luria–Nebraska neuropsychological battery: A case critique. *Journal of Consulting and Clinical Psychology*, 50, 32–39.
- Deutsch, S. I., & Davis, K. L. (1983). Schizophrenia: A review of diagnostic and biological issues: II. Biological issues. *Hospital and Community Psychiatry*, 34, 423–437.
- Drewe, E. A. (1975). An experimental investigation of Luria theory on the effect of frontal lobe lesions in man. *Neuropsychological*, 13, 421–429.
- Dunn, L. M., & Markwardt, F. C. (1970). *Peabody individual Achievement Test Manual*. Circle Pine, MN: American Guidance Service.
- Evans, D. A., Becken, L. A., Albert, M. S., Herbert, L. E., Schen, P. A., Funkenstein, H. H., & Taylor, J. O. (1993). Level of education and change in cognitive function in a community population of older persons. *Annals of Epidemiology*, 3, 71–77.
- Filskov, S. B., & Boll, T. J. (1981). *Handbook of clinical neuropsychology*. New York: Wiley-Interscience.
- Filskov, S. B., & Goldstein, S. G. (1974). Diagnostic validity of the Halstead–Reitan neuropsychological battery. *Journal of Consulting and Clinical Psychology*, 42, 382–388.
- Fitzhugh, K. B., Fitzhugh, L. C., & Reitan, R. M. (1961). Psychological deficits in relation to acuteness of brain dysfunction. *Journal of Consulting Psychology*, 25, 61–66.
- Fitzhugh, K. B., Fitzhugh, L. C., & Reitan, R. M. (1962). Wechsler–Bellevue comparisons in groups of 'chronic' and 'current' lateralized and diffuse brain lesions. *Journal of Consulting Psychology*, 26, 306–310.

- Fletcher, J. M., & Satz, P. (1980). Developmental changes in the neuropsychological correlates of reading achievement: A six-year longitudinal follow-up. *Journal of Clinical Neuropsychology*, 2, 23–37.
- Freedman, M. (1990). Parkinson's disease. In J. L. Cummings (Ed.), *Subcortical dementia* (pp. 108–122). New York: Oxford University Press.
- Goldberg, E., & Tucker, D. (1979). Motor preservation and long-term memory for visual forms. *Journal of Clinical Neuropsychology*, 1, 273–288.
- Golden, C. J. (1979). Identification of specific neurological disorders using double discrimination scales derived from the standardized Luria neuropsychological battery. *International Journal of Neuroscience*, 10, 51–56.
- Golden, C. J. (1980). In reply to Adams' "In search of Luria's battery: A false start.". *Journal of Consulting and Clinical Psychology*, 48, 517–521.
- Golden, C. J. (1981). A standardized version of Luria's neuropsychological tests: A quantitative and qualitative approach to neuropsychological evaluation. In S. B. Filskov, & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (pp. 608–642). New York: Wiley-Interscience.
- Golden, C. J., & Berg, R. A. (1983). Interpretation of the Luria–Nebraska neuropsychological battery by item intercorrelation: The memory scale. *Clinical Neuropsychology*, 5, 55–59.
- Golden, C. J., Gruber, B., Blose, L., Berg, R., Coffman, J., & Block, S. (1981). Differences in brain densities between chronic alcoholic and normal control patients. *Science*, 211, 508–510.
- Golden, C. J., Hammek, E. T., & Purisch, A. (1978). Diagnostic validity of the Luria neuropsychological battery. *Journal of Consulting and Clinical Psychology*, 46, 1258–1265.
- Golden, C. J., Hammek, T., & Purisch, A. (1980). *The Luria–Nebraska battery manual*. Los Angeles: Western Psychological Services.
- Golden, C. J., Moses, J. A., Zelazowski, R., Gruber, B., Zatz, L. M., Horvath, T. B., ... Berger, P. A. (1980). Cerebral ventricular size and neuropsychological impairment in young chronic schizophrenics. *Archives of General Psychiatry*, 37, 619–623.
- Golden, C. J., Purisch, A., & Hammek, T. (1985). *Luria–Nebraska neuropsychological battery manual—Forms I and II*. Los Angeles: Western Psychological Services.
- Goldstein, G. (1978). Cognitive and perceptual differences between schizophrenics and organics. *Schizophrenia Bulletin*, 4, 160–185.
- Goldstein, G. (1986a). The neuropsychology of schizophrenia. In I. Grant, & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (pp. 147–171). New York: Oxford University Press.
- Goldstein, G. (1986b). An overview of similarities and differences between the Halstead–Reitan and Luria–Nebraska batteries. In T. Incagnoli, G. Goldstein, & C. J. Golden (Eds.), *Clinical application of neuropsychological test batteries* (pp. 235–275). New York: Plenum Press.
- Goldstein, G. (1991). Comprehensive neuropsychological test batteries and research in schizophrenia. In S. R. Steinhauer, L. H. Gruzelier, & L. Zubin (Eds.), *Handbook of schizophrenia*. London: Elsevier.
- Goldstein, G. (1994). Cognitive heterogeneity in psychopathology: The case of schizophrenia. In P. Vernon (Ed.), *The neuropsychology of individual differences* (pp. 209–233). New York: Academic Press.
- Goldstein, G., & Beers, S. R. (1998). *Human brain function: Assessment and rehabilitation: Rehabilitation*. New York: Plenum Press.

- Goldstein, G., Beers, S. R., & Shemansky, W. L. (1996). Neuropsychological differences between schizophrenic patients with heterogeneous Wisconsin Card Sorting Test performance. *Schizophrenia Research*, 21, 1–18.
- Goldstein, G., Nussbaum, P. D., & Beers, S. R. (1998). *Human brain function: Assessment and rehabilitation: Neuropsychology*. New York: Plenum Press.
- Goldstein, G., & Ruthven, L. (1983). *Rehabilitation of the brain damaged adult*. New York: Plenum.
- Goldstein, G., & Shelly, C. (1971). Field dependence and cognitive, perceptual and motor skills in alcoholics: A factor analytic study. *Quarterly Journal of Studies on Alcohol*, 32, 29–40.
- Goldstein, G., & Shelly, C. (1972). Statistical and normative studies of the Halstead neuropsychological test battery relevant to a neuropsychiatric hospital setting. *Perceptual and Motor Skills*, 34, 603–620.
- Goldstein, G., & Shelly, C. H. (1975). Similarities and differences between psychological deficit in aging and brain damage. *Journal of Gerontology*, 30, 448–455.
- Goldstein, G., & Shelly, C. (1982). A further attempt to cross-validate the Russell, Neuringer, and Goldstein neuropsychological keys. *Journal of Consulting and Clinical Psychology*, 50, 721–726.
- Goldstein, G., & Shemansky, W. L. (1997). Patterns of performance by neuropsychiatric patients on the Halstead category test: Evidence for conceptual learning in schizophrenic patients. *Archives of Clinical Neuropsychology*, 12, 251–255.
- Goldstein, G., & Watson, J. R. (1989). Test-retest reliability of the Halstead-Reitan battery and the WAIS in a neuropsychiatric population. *The Clinical Neuropsychologist*, 3, 265–273.
- Goldstein, G., Shemansky, W. J., & Allen, D. N. (2005). Cognitive function in schizoaffective disorder and clinical subtypes of schizophrenia. *Archive of Clinical Neuropsychology*, 20(2), 153–159.
- Goldstein, K., & Scheerer, M. (1941). Abstract and concrete behavior: An experimental study with special tests. *Psychological Monographs*, 63. (Whole No. 239).
- Goodglass, H. (August, 1983). Aphasiology in the United States. In G. Goldstein (Chair), *Symposium: History of Clinical Neuropsychology in the United States, Ninety-first annual convention of the American Psychological Association*, Anaheim, CA.
- Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders* (2nd ed.). Philadelphia, PA: Lea & Febiger.
- Green, M. F., Nuechterlein, K. H., Gold, J. M., Barch, D. M., Cohen, J., Esscock, S., Fenton, W. S., Frese, F., Goldberg, T. E., Heaton, R. K., Keefe, R. S. E., Kern, R. S., Kraemer, H., Stover, E., Weinberger, D. R., Zalcman, S., & Marder, S. R. (2004). Approaching a consensus cognitive battery for clinical trials in schizophrenia: The NIMH-MATRICS conference to select cognitive domains and test criteria. *Biological Psychiatry*, 56, 301–307.
- Gruzelier, I.H. (1991). Hemispheric imbalance: Syndromes of schizophrenia, premorbid personality, and neurodevelopmental influences. In S. R. Steinhauer, J. H. Gruzelier, & J. Zubin (Eds.), *Handbook of schizophrenia: Vol. 5. Neuropsychology, Psychophysiology, and information processing* (pp. 599–650). Amsterdam: Elsevier.
- Guilmette, T. L., & Kastner, M. P. (1996). The prediction of vocational function from neuropsychological data. In R. L. Sbordone, & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 387–411). Delray Beach, FL: GR Press/St. Lucie Press.
- Halstead, W. C. (1947). *Brain and intelligence: A quantitative study of the frontal lobes*. Chicago: The University of Chicago Press.

- Harvey, P. D. (2014). What is the evidence for changes in cognition and functioning over the lifespan in patients with schizophrenia? *Journal of Clinical Psychiatry*, 75, 34–38.
- Heaton, R. K. (1980). *A manual for the Wisconsin Card Sorting Testing*. Odessa, FL: Psychological Assessment Resources, Inc.
- Heaton, R. K., Baade, L. E., & Johnson, K. L. (1978). Neuropsychological test results associated with psychiatric disorders in adults. *Psychological Bulletin*, 85, 141–162.
- Heaton, R. K., & Crowley, T. (1981). Effects of psychiatric disorders and their somatic treatment on neuropsychological test results. In S. B. Filskov, & T. J. Boll (Eds.), *Handbook of neuropsychology*. New York: Wiley-Interscience.
- Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan battery*. Odessa, FL: Psychological Assessment Resources.
- Heaton, R. K., & Pendleton, M. G. (1981). Use of neuropsychological tests to predict adult patients' everyday functioning. *Journal of Consulting and Clinical Psychology*, 49, 807–821.
- Henn, F. A., & Nasrallah, H. A. (1982). *Schizophrenia as a brain disease*. New York: Oxford University Press.
- Jarvis, P. E., & Barth, J. T. (1984). *Halstead-reitan test battery: An interpretive guide*. Odessa, FL: Psychological Assessment Resources.
- Jastak, S., & Wilkinson, G. S. (1984). *The wide range achievement test-revised*. Wilmington, DE: Jastak Associates, Inc.
- Jones, B. P., & Butters, N. (1983). Neuropsychological assessment. In M. Hersen, A. S. Bellack, & A. E. Kazdin (Eds.), *The clinical psychology handbook* (pp. 377–396). New York: Pergamon Press.
- Kaplan, E. (1979). Presidential address. Presented at the International Neuropsychological Society, Noordwijkerhout, Holland.
- Kaplan, E. H., Goodglass, H., & Weintraub, S. (1983). *The Boston naming test*. Philadelphia: Lea & Fibiger.
- Kertesz, A. (1979). *Aphasia and associated disorders: Taxonomy, localization and recovery*. New York: Grune & Stratton.
- Kimura, D. (1961). Some effects of temporal lobe damage on auditory perception. *Canadian Journal of Psychology*, 15, 156–165.
- Kinsbourne, M. (1980). Attentional dysfunctions and the elderly: Theoretical models and research perspectives. In L. W. Poon, J. L. Fozard, L. S. Cennak, D. Arenberg, & L. W. Thompson (Eds.), *New directions in memory and aging* (pp. 113–129). Hillsdale, NJ: Erlbaum.
- Kjove, H. (1974). Validation studies in adult clinical neuropsychology. In R. M. Reitan, & L. H. Davison (Eds.), *Clinical neuropsychology: Current status and applications* (pp. 211–235). Washington, DC: V.H. Winscon & Sons.
- Levin, H. S., Benton, A. L., & Grossman, R. G. (1982). *Neurobehavioral consequences of closed head injury*. New York: Oxford University Press.
- Lewis, R. F., & Rennick, P. M. (1979). *Manual for the repeatable cognitive-perceptual-motor battery*. Grosse Pointe Park, MI: Axon Publishing Co.
- Lezak, M. (1976). *Neuropsychological assessment*. New York: Oxford University Press.
- Lezak, M. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.
- Lyon, G. R., & Flynn, J. F. (1991). Educational validation studies with subtypes of learning-disabled readers. In B. P. Rourke (Ed.), *Neuropsychological validation of learning disability subtypes* (pp. 233–242). New York: The Guilford Press.
- Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books.

- Luria, A. R. (1973). *The working brain*. New York: Basic Books.
- Malec, J. (1978). Neuropsychological assessment of schizophrenia vs. brain damage: A review. *Journal of Nervous and Mental Disease*, 166, 507–516.
- Marcotte, T. D., & Grant, I. (2009). *Neuropsychology of everyday functioning*. New York: Guilford Press.
- Marsh, L., Lauriello, J., Sullivan, E. V., & Pfefferbaum, A. (1996). Neuroimaging in psychiatric disorders. In E. Bigler (Ed.), *Neuroimaging II: Clinical applications* (pp. 73–125). New York: Plenum Press.
- Matthews, C. G. (1981). Neuropsychology practice in hospital setting. In S. B. Filskov, & T. J. Boll (Eds.), *Handbook of clinical neuropsychology*. New York: Wiley-Interscience.
- McCue, M. (1997). The relationship between neuropsychology and functional assessment in the elderly. In P. D. Nussbaum (Ed.), *Handbook of neuropsychology and aging* (pp. 394–408). New York: Plenum Press.
- McCue, M., Shelly, C., Goldtein, G., & Katz-Garris, L. (1984). Neuropsychological aspects of learning disability in young adults. *Clinical Neuropsychology*, 6, 229–233.
- McKay, S., & Golden, C. J. (1979a). Empirical derivation of experimental scales for the lateralization of brain damage using the Luria–Nebraska neuropsychological battery. *Clinical Neuropsychology*, 1, 1–5.
- McKay, S., & Golden, C. J. (1979b). Empirical derivation of experimental scales for localized brain lesions using the Luria–Nebraska neuropsychological battery. *Clinical Neuropsychology*, 1, 19–23.
- McKay, S. E., & Golden, C. J. (1981). The assessment of specific neuropsychological skills using scales derived from factor analysis of the Luria–Nebraska neuropsychological battery. *International Journal of Neuroscience*, 14, 189–204.
- Meier, M. J. (1974). Some challenges for clinical neuropsychology. In R. M. Reitan, & L. A. Davison (Eds.), *Clinical neuropsychology: Current status and applications* (pp. 289–323). Washington, DC: V.H. Winston and Sons.
- Meier, M. J., Benton, A. L., & Diller, L. (1987). *Neuropsychological rehabilitation*. Edinburgh: Churchill Livingstone.
- Meyers, J. E., & Rohling, M. L. (2004). Validation of the Meyers short battery on mild TBI patients. *Archives of Clinical Neuropsychology*, 19, 637–651.
- Minshew, N. J., Goldstein, G., Dombrowski, S. N., Panchalingam, K., & Petlegrew, J. W. (1993). A preliminary 31 p-NMR study of autism: Evidence for under synthesis and increased degradation of brain membranes. *Biological Psychiatry*, 33, 762–773.
- Mirsky, A. F., Anthony, B. J., Duncan, C. C., Ahearn, M. B., & Kellam, S. G. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review*, 2, 109–145.
- Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology*, 2, 219–226.
- Moses, J. A., Golden, C. J., Berger, P. A., & Wisniewski, M. (1981). Neuropsychological deficits in early, middle, and late stage Hunting's disease as measured by the Luria–Nebraska neuropsychological battery. *International Journal of Neuroscience*, 14, 95–100.
- Moses, J. A., Jr., & Purisch, A. D. (1997). The evolution of the Luria–Nebraska battery. In G. Goldstein, & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 131–170). New York: Plenum Press.
- Newcombe, F. (1969). *Missile words of the brain: A study of psychological deficits*. Oxford: The Clarendon Press.

- Newman, O. S., Heaton, R. K., & Lehman, R. A. W. (1978). Neuropsychological and MMPI correlates of patients' future employment characteristics. *Perceptual and Motor Skills*, 46, 635–642.
- Nussbaum, P. D. (Ed.), (1997). *Handbook of neuropsychology and aging*. New York: Plenum Press.
- Osterrieth, P. A. (1944). Filetest de copie d'une figure complexe: Contribution à l'étude de la perception et de la mémoire [The test of copying a complex figure: A contribution to the study of perception and memory]. *Archives de Psychologie*, 30, 286–356.
- Palmer, B. W., Heaton, R. K., Paulsen, J. S., Kuck, J., Braff, D., HruTis, M. J., ... Jeste, D. Y. (1997). Is it possible to be schizophrenic yet neuropsychologically normal? *Neuropsychology*, 11, 437–446.
- Purisch, A. D., Golden, C. J., & Hammeke, T. A. (1978). Discrimination of schizophrenic and brain-injured patients by a standardized version of Luria's neuropsychological tests. *Journal of Consulting and Clinical Psychology*, 46, 1266–1273.
- Purisch, A. D., & Sbordone, R. J. (1986). The Luria–Nebraska Neuropsychological Battery. In G. Goldstein, & R. E. Tarter (Eds.), *Advances in Clinical Neuropsychology* (Vol. 3). New York: Plenum Press.
- Randolph, C. (1998). *RBANS manual. Repeatable Battery for Assessment of Neuropsychological Status*. San Antonio, TX: The Psychological Corporation.
- Ragland, J. D., Ranganath, C., Phillips, J., Boudewyn, M. A., Kring, A. M., Lesh, T. A., & Carter, C. S. (2015). Cognitive control of episodic memory in schizophrenia: Differential role of dorsolateral and ventrolateral prefrontal cortex. *Frontiers in Human Neuroscience*, 9(604). Available from <https://doi.org/10.3389/fnhum.2015.00604>.
- Reed, J. (August, 1983). The Chicago–Indianapolis group. In G. Goldstein (Chair). *Symposium: History of human neuropsychology in the United States. Ninety-first annual convention of the American Psychological Association*, Anaheim, CA.
- Reed, J. C., & Reed, H. B. C. (1997). The Halstead–Reitan neuropsychological battery. In G. Goldstein, & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 93–129). New York: Plenum Press.
- Reitan, R. M. (1955). An investigation of the validity of Halstead's measures of biological intelligence. *Archives of Neurology and Psychiatry*, 73, 28–35.
- Reitan, R. M. (1958). Qualitative versus quantitative mental changes following brain damage. *The Journal of Psychology*, 46, 339–346.
- Reitan, R. M. (1959). Correlations between the trail making test and the Wechsler–Bellevue scale. *Perceptual and Motor Skills*, 9, 127–130.
- Reitan, R. M. (1964). Psychological deficits resulting from cerebral lesions in man. In J. M. Warren, & K. Aken (Eds.), *The frontal granular cortex and behavior* (pp. 295–312). New York: McGraw-Hill.
- Reitan, R. M. (1966). A research program on the psychological effects of brain lesions in human beings. In N. R. Ellis (Ed.), *International review of research in mental retardation* (pp. 153–218). New York: Academic Press.
- Reitan, R. M. (August, 1973). Behavioral manifestations of impaired brain functions in aging. In J. L. Fozard (Chair), *Similarities and differences of brain–behavior relationships in aging and cerebral pathology*. Symposium presented at the American Psychological Association, Montreal, Canada.
- Reitan, R. M. (1987). *The neuropsychological deficit scale for adults. Computer program*. Tucson, AZ: Neuropsychology Press.
- Reitan, R. M. (1991). *The Neuropsychological deficit scale for adults. Users program*. Tucson, AZ: Neuropsychology Press.

- Reitan, R. M., Hom, J., & Wolfson, D. (1988). Verbal processing by the brain. *Journal of Clinical and Experimental Neuropsychology, 10*, 400–408.
- Reitan, R. M., & Wolfson, D. (1993). *The Halstead–Reitan neuropsychological test battery: Theory and clinical interpretation* (2nd ed.). Tucson: Neuropsychology Press.
- Reitan, R. M., & Wolfson, D. (1995). Influence of age and education on neuropsychological test results. *The Clinical Neuropsychologist, 9*, 151–158.
- Reitan, R. M., & Wolfson, D. (1997). Emotional disturbances and their interaction with neuropsychological deficits. *Neuropsychology Review, 7*, 3–19.
- Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives de Psychologie, 28*, 286–340.
- Rourke, B. P. (Ed.), (1985). *Neuropsychology of learning disabilities: Essentials of subtype analysis*. New York: The Guilford Press.
- Russell, E.W. (1993). *Halstead–Russell neuropsychological evaluation system, norms and conversion tables* (unpublished manuscript).
- Russell, E. W. (1975a). A multiple scoring method for the assessment of complex memory functions. *Journal of Consulting and Clinical Psychology, 43*, 800–809.
- Russell, E. W. (1975b). Validation of a brain damage versus schizophrenia MMPI. *Journal of Clinical Psychology, 33*, 190–193.
- Russell, E. W. (1981). The pathology and clinical examination of memory. In S. B. Filskov, & T. J. Boll (Eds.), *Handbook of clinical neuropsychology*. New York: Wiley-Interscience.
- Russell, E. W. (1997). Developments in the psychometric foundations of neuropsychological assessment. In G. Goldstein, & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 15–65). New York: Plenum.
- Russell, E. W., & Levy, M. (1987). Revision of the Halstead category test. *Journal of Consulting and Clinical Psychology, 55*, 898–901.
- Russell, E. W., Neuringer, C., & Goldstein, G. (1970). *Assessment of brain damage: A neuropsychological key approach*. New York: Wiley-Interscience.
- Ryan, C. M. (1998). Assessing medically ill patients: *Diabetes mellitus* as a model disease. In G. Goldstein, P. D. Nussbaum, & S. R. Beers (Eds.), *Human brain function: Assessment and rehabilitation: Neuropsychology* (pp. 227–245). New York: Plenum Press.
- Satz, P., Taylor, H. G., Friel, J., & Fletcher, J. M. (1978). Some developments and predictive precursors of reading disability. In A. L. Benton, & D. Pearl (Eds.), *Dyslexia: An appraisal of current knowledge* (pp. 313–347). New York: Oxford University Press.
- Sbordone, R. J., & Long, C. J. (Eds.), (1996). *Ecological validity of neuropsychological testing*. Delray Beach, FL: GR Press/St. Lucie Press.
- Schear, J. M. (1984). Neuropsychological assessment of the elderly in clinical practice. In P. E. Logue, & J. M. Schear (Eds.), *Clinical neuropsychology: A multidisciplinary approach* (pp. 199–235). Springfield, IL: C. C. Thomas.
- Schear, J. M. (1987). Utility of cluster analysis in classification of mixed neuropsychiatric patients. *Archives of Clinical Neuropsychology, 2*, 329–341.
- Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Nerosurgery, and Psychiatry, 20*, 11–21.
- Seashore, C. B., Lewis, C., & Saaeteweit, J. G. (1960). *Seashore measures of musical talent: Manual*. San Antonio, TX: Psychological Corporation.
- Selz, M., & Reitan, R. M. (1979). Rules for neuropsychological diagnosis: Classification of brain function in older children. *Journal of Consulting and Clinical Psychology, 47*, 258–264.

- Semmes, J., Weinstein, S., Ghent, L., & Teuber, H.-L. (1960). *Somatosensory changes after penetrating brain wounds in man*. Cambridge, MA: Harvard University.
- Shaw, D. (1966). The reliability and validity of the Halstead category test. *Journal of Clinical Psychology*, 22, 176–180.
- Shelly, C., & Goldstein, G. (1983). Discrimination of chronic schizophrenia and brain damage with the Luria–Nebraska battery: A partially successful replication. *Clinical Neuropsychology*, 5, 82–85.
- Sherrill, R. E., Jr. (1987). Options for shonening Halstead's category test for adults. *Archives of Clinical Neuropsychology*, 5, 82–85.
- Smith, A. (1965). Cenain hypothesized hemispheric differences in language and visual functions in human adults. *Cortex*, 2, 109–126.
- Smith, A. (1966a). Intellectual functions in patients with lateralized frontal tumors. *Journal of Neurology, Neurosurgery, and Psychiatry*, 29, 52–59.
- Smith, A. (1966b). Verbal and nonverbal test performance of patients with 'acute' lateralized brain lesions (tumors). *Journal of Nervous and Mental Disease*, 141, 517–523.
- Smith, A. (1975). Neuropsychological testing in neurological disorders. In W. J. Friedlander (Ed.), *Advances in neurology* (Vol. 7, pp. 49–110). New York: Raven Press.
- Sperry, R. W., Gazzaniga, M. S., & Bogen, J. E. (1969). Interhemispheric relationships: The neocortical commissures; syndromes of hemisphere disconnection. In P. J. Vinkin, & G. W. Bruyn (Eds.), *Handbook of clinical neurology*. Amsterdam: North Holland.
- 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.
- Spreen, O., & Strauss, E. (1988). *A compendium of neuropsychological tests* (2nd ed.). New York: Oxford University Press.
- Stambrook, M. (1983). The Luria–Nebraska neuropsychological battery: A promise that may be partly fulfilled. *Journal of Clinical Neuropsychology*, 5, 247–269.
- Stein, D. G. (1988). Contextual factors in recovery from brain damage. In A.-L. Christensen, & B. P. Uzzell (Eds.), *Neuropsychological rehabilitation* (pp. 1–18). Boston: Kluwer Academic Press.
- Steinhauer, S. R., Hill, S. Y., & Zubin, J. (1987). Event-related potentials in alcoholics and their first-degree relatives. *Alcohol*, 4, 307–314.
- Stern, R. A., & White, T. (2003). *Neuropsychological Assessment Battery (NAB)*. Lutz FL: Psychological Assessment Resources (PAR).
- Swiercinsky, D. (1978). *Manual for the adult neuropsychological evaluation*. Springfield, IL: C. C. Thomas.
- Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness: A practical scale. *Lancet*, 2, 81–84.
- Teuber, H.-L. (1959). Some alterations in behavior after cerebral lesions in man. In A. D. Bass (Ed.), *Evolution of nervous control from primitive organisms to man* (pp. 157–194). Washington, DC: American Association for Advancement of Science.
- Teuber, H.-L. (1964). The riddle of frontal lobe function in man. In J. M. Warren, & K. Akert (Eds.), *The frontal granular cortex and behavior* (pp. 410–441). New York: McGraw-Hill.
- Teuber, H.-L., Battersby, W. S., & Bender, M. B. (1951). Performance of complex visual tasks after cerebral lesions. *The Journal of Nervous and Mental Disease*, 114, 413–429.
- Teuber, H.-L., & Weinstein, S. (1954). Performance on a form-board task after penetrating brain injury. *Journal of Psychology*, 38, 177–190.
- Townes, B. D., Martin, D. C., Nelson, D., Prosser, Pepping, M., Maxwell, J., . . . Prestol, M. (1985). Neurobehavioral approach to classification of psychiatric patients using a competency model. *Journal of Consulting and Clinical Psychology*, 53, 33–42.

- Vega, A., & Parsons, O. (1967). Cross-validation of the Halstead–Reitan tests for brain damage. *Journal of Consulting Psychology*, 31, 619–625.
- Warrington, E. K., & James, M. (1991). *Visual object and space perception battery*. Suffolk, England/Gaylord, ML: Thames Valley Test Co./National Rehabilitation Services.
- Wechsler, D. (1945). *Wechsler Memory Scale Manual*. New York: Psychological Corporation.
- Wechsler, D. (1987a). *Wechsler Adult Intelligence Scale-Revised*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1987b). *Wechsler Memory Scale-Revised*. New York: Psychological Corporation.
- Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale III*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (2009). *Wechsler Memory Scale IV*. San Antonio, TX: Pearson.
- Wechsler, D. (1997b). *Wechsler Memory Scale III*. San Antonio, TX: The Psychological Corporation.
- Werthheimer, M. (1923). Studies in the theory of Gestalt psychology. *Psychologische Forschung*, 4, 301–350.
- Whearty, K. M., Allen, D. N., Lee, B. G., & Strauss, G. P. (2015). The evaluation of insufficient cognitive effort in schizophrenia in light of low IQ scores. *Journal of Psychiatric Research*, 68, 397–404.
- Woodward, N. D. (2016). The course of neuropsychological impairment and brain structure abnormalities in psychotic disorders. *Neuroscience Research*, 102, 39–46.
- Yozawitz, A. (1986). Applied neuropsychology in a psychiatric center. In I. Grant, & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (pp. 121–146). New York: Oxford University Press.
- Zimmer, B., & Grossberg, G. (1997). Geriatric psychopharmacology: An update and review. In P. D. Nussbaum (Ed.), *Handbook of neuropsychology* (pp. 483–507). New York: Plenum Press.

## Further reading

- Golden, C. J. (1978). *Diagnosis and rehabilitation in clinical neuropsychology*. C.C. Thomas: Springfield, IL.
- Jeannerod, M. (Ed.). (1987). *Neurophysiological and neuropsychological aspects of spatial neglect*. North Holland: Amsterdam.
- Luria, A. R. (1970). *Traumatic aphasia*. The Hague: Mouton and Co. Printers.
- Squire, L. R., & Butters, N. (Eds.). (1984). *Neuropsychology of memory*. New York: The Guilford Press.

# Assessment in sports: psychological and neuropsychological approaches

9

Ruben J. Echemendia<sup>1,2</sup>, Frank M. Webbe<sup>3</sup>, Victoria C. Merritt<sup>4</sup> and  
Gabriela González<sup>3</sup>

<sup>1</sup>UOC Concussion Care Clinic, State College, Pennsylvania, PA, United States, <sup>2</sup>University of Missouri-Kansas City, Pennsylvania, PA, United States, <sup>3</sup>Florida Institute of Technology, Melbourne, FL, United States, <sup>4</sup>VA San Diego Healthcare System, San Diego, CA, United States

## Psychological assessment in sport

Perhaps one of the most sought out endeavors within the athletic community involves identifying variables that foster enhanced athletic performance. Usain Bolt, Simone Biles, Michael Phelps, and Kerri Walsh are arguably just a few of the world's top athletes, who have achieved the zenith of their particular sports. However, one cannot help but wonder, what specific variables contribute to such remarkable and elite sports performances? Obviously, physical attributes including flexibility, agility, speed, stamina, and strength have a significant impact on athletic prowess. Years of hard work and targeted training are also vital contributors to peak level performance. Others might even add that environmental factors such as accessibility to gyms, socioeconomic status, family support, and quality of sports programs and coaching promote enhanced athletic performance. Still, should athletic success be solely attributed to raw physical talent and 10,000 hours of targeted training? Are there other influential underlying psychological commonalities among these stellar athletes that positively impact their athletic achievement? When Yogi Berra quipped, "baseball is 90% mental and the other half is physical" there was more than humor in the statement. At elite levels of sport, many athletes share great native ability but not all share great mental control. Some of these underlying mechanisms may include psychological factors such as confidence, competitiveness, aggressiveness, and determination. If such psychological mechanisms affect athletic achievement, how can we measure and further develop such psychological skills? Can these measures be used to predict potential athletic achievement? Are there specific psychological mechanisms that influence an athlete's level of success within their sport? These questions will be addressed in the first portion of this chapter.

## Disciplines of sports psychology

The Society for Sport, Exercise, and Performance Psychology (Division 47 of the American Psychological Association) defines sport psychology thusly: “Sport psychology is a proficiency that uses psychological knowledge and skills to address optimal performance and well-being of athletes, developmental and social aspects of sports participation, and systemic issues associated with sports settings and organizations” ([APA, 2009](#)). In their popular text, [Weinberg and Gould \(2014\)](#) define sport psychology more simply as “The scientific study of people and their behaviors in sport and exercise activities and the practical application of that knowledge.” Other definitions abound but these two capture the core of most. It is the application of empirically validated approaches in program development and intervention that differentiate sport psychology from other approaches that aim to improve athletic performance.

Although it may appear that public fascination with all things sport is a somewhat recent phenomenon, the psychological study of sport extends back more than 125 years. Credit for the earliest systematic study of the psychology of sport accrues to Indiana University psychologist Norman Triplett. Beginning in the mid 1890s he initiated a series of laboratory experiments examining social-facilitation effects in “wheelmen” (bicyclists; [Triplett, 1898](#)). By varying the presence and absence of others whom he referred to as pacers, Triplett concluded that the presence of others generally resulted in increases in speed over successive trials. Triplett amassed extensive data showing these effects in children and adults of both sexes. Given that historians date the origin of scientific psychology to Wilhelm Wundt’s publication of his *Grundzüge der Physiologischen Psychologie* in 1873 and his founding of a laboratory of experimental psychology at Leipzig 6 years later ([Boring, 1929](#)), sport psychology was close to being a founding disciplinary area of psychology itself. In modern sport psychology, multiple disciplines and domains exist including: biomechanics, exercise physiology, motor development, motor learning, performance enhancement, and sport sociology. As will be developed later in this chapter, the merging of neuropsychology into the realm of sport—and the resultant hybrid discipline—*sport neuropsychology*, has drawn considerable attention with the study of sport-related concussions (SRC).

Neuropsychology extends back even further than sport psychology, at the very least to Paul Broca’s study in 1861 of the brain origins of the speech difficulties exhibited by the aphasic patient LeBorgne (named “Tan” in Broca’s original paper). Upon LeBorgne’s death, Broca’s autopsy confirmed that a stroke had destroyed a significant amount of neural tissue in the posterior frontal lobe—the section now known as Broca’s area ([Broca, 1861](#)). However, it has been suggested that neuropsychology as a specific discipline really did not differentiate itself from neurology and psychology until the decade of the 1940s ([Lezak, 1983](#)).

Modern neuropsychology studies the relationship between the functioning brain and behavior, with behavior often broken down into domains that include intellectual, language, processing, executive control, learning, memory, visuospatial, and emotional. Both sport psychology and neuropsychology share several similarities, a

primary one being that each has its scientific and its applied sides. Experimental neuropsychology uses methods from experimental psychology to uncover the relationship between the nervous system and cognitive function, with both human and animal studies reported. Although experimental neuropsychology was the original parent discipline, clinical neuropsychology predominates today, and experimental neuropsychology has become more recognized as part of the discipline known as cognitive neuroscience. In describing the recognized proficiency of clinical neuropsychology, the American Psychological Association defines the discipline in this way:

*Clinical Neuropsychology is a specialty in professional psychology that applies principles of assessment and intervention based upon the scientific study of human behavior as it relates to normal and abnormal functioning of the central nervous system. The specialty is dedicated to enhancing the understanding of brain–behavior relationships and the application of such knowledge to human problems.*

[APA \(2010\).](#)

In sport psychology, exercise science is the predominant scientific side, but the psychology half also is divided into clinical versus scientific aspects. Obvious areas of overlapping interest exist between sport psychology and neuropsychology. For example, exercise science studies motor control and motor learning in sport. Brain injuries might obviously impact such learning and performance, and the rehabilitative effects of re-learning motoric behavior might in turn affect recovery processes in the brain. A sport neuropsychological approach would map such relationships. Another study might examine the role of excessive metabolic demands in endurance sports in altering brain function and cognitive performance. One of the more amazing *tours de force* in this regard was rudimentary neuropsychological testing of noted mountain climber and film producer David Breashears as he and his climbing partner, Ed Viesturs, “relaxed” on the summit of Mount Everest with no supplemental oxygen ([Public Broadcasting Service, 1997](#)). Remarkably, his performance was better than many people who are at sea level.

## **Unique aspects of the sport milieu**

The obvious commonality of both Sport Psychology and Sport Neuropsychology is the actual sport setting, and the methodological, professional, and ethical complexities of the setting that provide a rich area for scientific study while also taxing the practitioner and researcher alike. Among the challenges presented by the sports arena are ethical concerns in both science and practice. [Bush and Iverson \(2011\)](#) have identified seven areas in sport neuropsychology where ethical questions are likely to arise: (1) conflicts of interest in research and program development; (2) competence; (3) bases for scientific and professional judgments; (4) relationships and roles; (5) test selection, use, and interpretation; (6) privacy, confidentiality, and informed consent; and (7) avoidance of false or deceptive advertising. These same considerations apply to sport psychology as well.

One of the most common ethical issues that arises in both sport psychology and sport neuropsychology concerns identifying the client (Echemendia & Bauer, 2015). In a typical clinical psychology setting, the client engages the practitioner, and there is no confusion about the relationship. In sport psychology and sport neuropsychology clients may be contemporaneous, such as when a psychologist works for a team or organization but concurrently has the task of working with individual athletes. The interests of the individual athlete may be different than the interests of the team, which may place the practitioner in an ethical quandary. It is incumbent upon the psychologist to clarify roles and relationships at the outset, before the organization and/or the individual athletes engage the practitioner's services. The obvious follow-up to delineating roles and relationships is to clarify any limits on privacy and confidentiality of information gleaned in the activity and the extent to which individual athlete information may be made available to the team and vice versa (Echemendia & Bauer, 2015). Moran (2012) also has warned sport psychologists to be aware of pitfalls inherent to the medical model. For example, he cautioned that a medical model often creates the expectation of a quick cure or fix for a problem. Such optimism may not be warranted in many circumstances. A second issue that is common but generally not discussed is that the client—whether athlete or coach—has a level of expertise that may rival the psychologist. Consider that a professional golfer may engage a swing coach to assist in ridding his swing of flaws. The golfer obviously has superior knowledge of the essentials of the golf swing and the mental processes involved, perhaps more than the coach. The coach understands that her role is to collaborate with the golfer to identify flaws and make recommendations for overcoming them. In most instances, sport psychologists will know less about sport-specific behaviors than the athlete client, who has spent most of a lifetime engaged in that sport. Credibility must be earned by the practitioner's psychological expertise, not his or her sports expertise. *However*, the would-be sport psychologist or sports neuropsychologist who walks into a locker room with little or no sport-specific knowledge will likely fail since it is clear to the athletes that this person has done little background preparation. Unlike many clinical psychology clients, athletes are likely to be incredibly motivated to improve, and are generally not hesitant to work diligently and almost obsessively if they have confidence in the psychologist and the intervention (Murphy, 1995).

## **Approaches to assessment in sport psychology**

It is common knowledge that assessments can be approached in various forms, each with limitations and advantages. However, a key issue in assessment for both sports psychology and sports neuropsychology involves individual versus group formats for testing. Sports teams vary in size from relatively small teams (e.g., basketball) to very large squads such as football. The psychologist is often working with more than one team in a league or school. Given the sheer numbers of individuals involved, it is often impractical and too costly to assess individual athletes in a

one-on-one setting. Each approach has clear strengths and limitations. Disadvantages to group testing include:

1. *Distractibility*: An athlete's attention may be diverted from the task at hand due to environmental stimuli such as rowdy or talkative teammates.
2. *Behavioral observations*: In a sense, the primary and most important responsibility for someone proctoring a group test administration would be to maintain order and silence in the testing room, which hinders a proctor's ability to identify relevant individual behavioral observations that could impact test results and interpretation. For example, a proctor may be busy silencing two athletes and miss another athlete who may be demonstrating poor concentration or fatigue, or misinterpreting test instructions.
3. *Rapport*: Rapport refers to the ability for two or more individuals to relate in an honest and forthcoming manner. This type of relationship is important in this scenario since athletes are more likely to respond truthfully to test items if they believe the practitioner has their best interests at heart. Within the group setting, establishing rapport may be challenging as there are far more athletes than clinicians, which minimizes the opportunity for individualized attention.
4. *Social desirability bias* ([Edwards, 1957](#)): Like most individuals, athletes may feel pressured to answer test items according to social or, in this case, team norms to avoid embarrassment or seeming incompetent. Thus, athletes may likely project a more favorable image than is actually the case to their teammates and coaches.

Practitioners and researchers can help mitigate some of these disadvantages. A few considerations they may bear in mind include providing noise canceling devices and/or workspace dividers to reduce the potential for distraction. Furthermore, practitioners are encouraged to be mindful of the testing room layout and decorations; specifically, individuals completing assessments may benefit from minimal wall decor and organized desks lined facing forward. White noise machines are another resourceful device to help mitigate distractibility.

With regards to rapport building, practitioners are encouraged to clearly and reasonably explain testing objectives, procedures, and behavioral expectations. Full disclosure of their credentials and the purposes for testing is also encouraged. Before commencing test administration, practitioners and researchers should attend to immediate concerns or questions in a genuine manner.

There are also advantages to group testing. First, group testing provides increased efficiency in test administration compared to the more time consuming individual testing. The time to administer psychological/personality measures may vary from 30 to 120 min. If every player on a football team were to be assessed individually (average 105 athletes), this would mean that a practitioner would spend at least  $\sim 55$  h to test the team compared to five 30–45 min sessions if tested in a group setting. However, as will be discussed later, the optimal number of athletes in a group testing situation may be 5 or less. Second, individualized testing involves a greater number of test administrators, which requires significantly greater financial and personnel resources than that required in group testing.

On the other hand, there are advantages to individual testing formats, which are largely inversely related to the disadvantages of group testing. For instance, an athlete is more likely to receive individual one-on-one interaction, which may in turn

**Table 9.1** Important considerations when administering assessment instruments in the sport setting

Disregarding the purpose of the instrument
Assuming that an inventory can be used across multiple settings
Using trait (fixed) inventories to study an athlete's state (transient) thoughts and behaviors
Disregarding the age ranges for which the instrument was designed
Overreliance on the consultant
Overreliance on test results rather than behavioral observations or interview data
Failure to establish rapport
Disregarding cultural factors and group differences
Failure to take into consideration test taker's attitude towards the process

increase test validity, clinician-client rapport, and gathering of observational/qualitative data. Athletes will most likely be less distracted by their teammates or concerned with how their responses may portray them, and the social desirability phenomena may decrease.

In order to increase the chances of obtaining valid results, test administrators must strive to provide optimal testing conditions for the athletes. Regardless of the testing approach (individual vs group) testing conditions should include:

1. appropriate lighting,
2. spacious room,
3. desks or tables lined facing forward,
4. noise control (may provide noise canceling devices, white noise machine, sound proof walls),
5. minimal distractors (e.g., distracting décor),
6. cleanliness and organization.

[Brinthaupt and Anshel \(2015\)](#) have recommended that practitioners take certain precautions before, during, and after psychological test administration in the sport setting (see **Table 9.1**). Although many of these recommendations apply broadly to any testing situation, some are directly relevant to the sport environment. Prior to testing, they suggest that practitioners identify the assessment's purpose, objective, and psychometric properties. Practitioners should also establish how test data will be handled and obtain the necessary consents and approvals. During and after test administration, researchers have advised that practitioners establish rapport with the athletes, take into consideration cultural and/or group differences, identify reading level of athletes, ensure data accuracy, interpret scores using appropriate norms, and document the experience.

## Most frequent constructs and behaviors measured in sports assessments

[Table 9.2](#) lists nine constructs that are frequently studied in sport psychology and common tests that sport psychologists employ to measure them. These tests and

**Table 9.2** Common constructs and behaviors measured in sport psychology psychological inventories

Personality	NEO-PI-3, NEO-FFI-3 (McCrae, Costa, & Martin, 2005)
Motivation	Sport Motivation Scale-6 (SMS-6; Mallett, Kawabata, Newcombe, Otero-Forero, & Jackson, 2007); Exercise Motivation Scale (EMS; Li, 1999)
Mental toughness	The Mental Toughness Scale (MTS; Madrigal et al., 2013)
Emotion regulation	Psychological Performance Inventory (PPI; Loehr, 1986)
Coping skills	Profile of Mood States (McNair, Lorr, & Droppleman, 1971); current version is POMS-2 (Heuchert & McNair, 2012).
Resiliency	Coping inventory for stressful situations (CISS) by Endler and Parker (1999); The Coping Style in Sport Survey (1990) by Anshel (1990)
Burn out/ overtraining	Connor—Davidson resilience scale (CD-RISC; Connor & Davidson, 2003)
Team cohesion	Connor—Davidson resilience scale (CD-RISC; Connor & Davidson, 2003)
Goal setting	Athlete burnout questionnaire (ABQ; Raedeke & Smith, 2001)
Mental health	The Sports Cohesiveness Questionnaire (Martens, Landers, & Loy, 1972); Multidimensional Sport Cohesion Instrument (Yukelson, Weinberg, & Jackson, 1984)
	Test of Performance Strategies (TPS; Thomas, Murphy, & Hardy, 1999)
	Assorted screening instruments in the NCAA's Mental Health Best Practices ( <a href="https://www.ncaa.org/sites/default/files/HS_Mental-Health-Best-Practices_20160317.pdf">https://www.ncaa.org/sites/default/files/HS_Mental-Health-Best-Practices_20160317.pdf</a> )

Source: Adapted from Brinthaupt, T. M., & Anshel, M. H. (2015). Practical guidelines for using inventories in sport psychology. *The Sport and Exercise Scientist*, 45, 12–13.

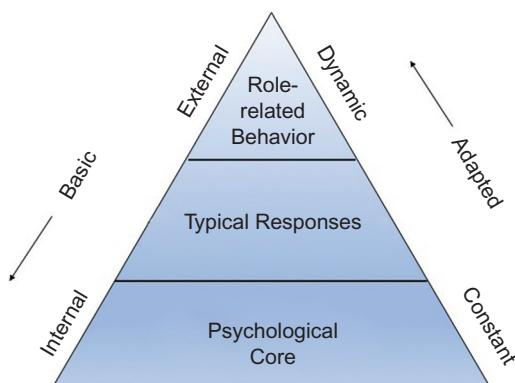
constructs are by no means all-inclusive since there are many other constructs and hundreds of tests used to measure them. This listing is meant only as an example. Indeed, in sport psychology there is continuing evaluation and evolution both of constructs and the scales that purport to measure them. Mental toughness is a good example since it has proven to be a tricky variable to measure with any degree of reliability and validity (Guccardi & Gordon, 2011). Indeed, conceptually, mental toughness crosses over into the resilience construct, coping, and mental health.

Test developers and researchers have studied certain psychological constructs that may interfere or enhance an athlete's ability to perform at the highest level. There is no one single psychological assessment that can predict enhanced sport performance; thus, the use of multiple psychological measures to obtain a broader and reliable picture of an athlete's internal mechanisms and potential capabilities may be more appropriate (Webbe, Salinas, Tiedemann, & Quackenbush, 2011). These results could then be translated into performance potential. It is important to note, that psychological constructs alone cannot predict elite performance, and those professionals who are interpreting test data are encouraged to examine other physical, environmental, and cultural factors that may contribute to the prediction model.

## Personality

Researchers, coaches, and team managers have frequently pursued the assessment of athletes' personality traits in hope that certain characteristics will relate to success in sport. Within sport psychology, personality is commonly defined as a set of relatively exclusive and unchanging core traits and tendencies that are possibly influenced by the environment and define an individual (Webbe et al., 2011). Martens (1975) described personality as the interaction between an individual's internal and constant Psychological Core (bottom tier), their Typical Responses (middle), and external and dynamic Role-related Behaviors (top) in the elaboration of personality (see Fig. 9.1 below). The Psychological Core is considered to be the most basic level of an individual's personality and consists of the individual's personal values, motives, and beliefs. Typical Responses refer to the usual ways in which individuals respond to external stimuli and Role-related Behaviors are dynamic and adapt to the individual's perception of the environment (Martens, 1975).

Measuring personality traits or characteristics in sport psychology generally assumes that the individual is healthy, which suggests use of scales that are oriented to healthy as opposed to abnormal functioning. Thus, it is relatively rare to see measures of psychopathology such as the Minnesota Multiphasic Personality Inventory (MMPI; any version) employed. More common are the NEO-PI (NEO-PI-R; Costa & McCrae, 1992) and the shorter NEO-FFI (NEO-FFI; Costa & McCrae, 1992), the 16PF (16PF; Cattell, Cattell, & Cattell, 1993), and occasionally the Personality Assessment Inventory (PAI; Morey, 2007). For example, Nia and Besharat (2010) compared athlete characteristics in individual and team sports using the NEO-PI-R and concluded that athletes who engaged in team sports scored significantly higher on Agreeableness as opposed to the individual sport athletes who evidenced higher scores on Conscientiousness. In sport psychology, differences are often found both between groups and among individuals within each group. The meaning of such



**Figure 9.1** Martens' model of personality structure.

differences is frequently not clear since the differences are generally not aligned with an overriding theory. Perhaps due to this lack of guiding theories, it is fair to say that the study of personality traits in the context of predicting success in sport has not produced an impressive body of literature (Webbe et al., 2011).

## Emotion regulation

Over the past 25 years, the Profile of Mood States (POMS) has been identified as one of the most widely used and accepted measures of mood within the sport and exercise setting (Leunes & Burger, 2000). Past research suggests that elite athletes who exhibit low levels of the negative mood states, anxiety, anger, depression, confusion, and fatigue, and increased levels of vigor—as presented in the POMS—will evidence greater athletic success than those who do not (Gutmann, Pollock, Foster, & Schmidt, 1984; McNair, Lorr, & Dropplerman, 1971; Silva, Shultz, Haslam, Martin, & Murray, 1985). Fig. 9.2 presents the “Iceberg” profile, which is characterized by a high score on the Vigor scale with lower scores on the other domains. This “Iceberg profile,” first identified by Morgan and Johnson (1978), is a commonly referenced profile that has been used with athletes across many sports and levels. Mainwaring (2011) and colleagues have shown that the iceberg frequently flattens or inverts following concussion and represents a reliable marker of concussion as expressed in mood alterations from baseline levels among some individuals.

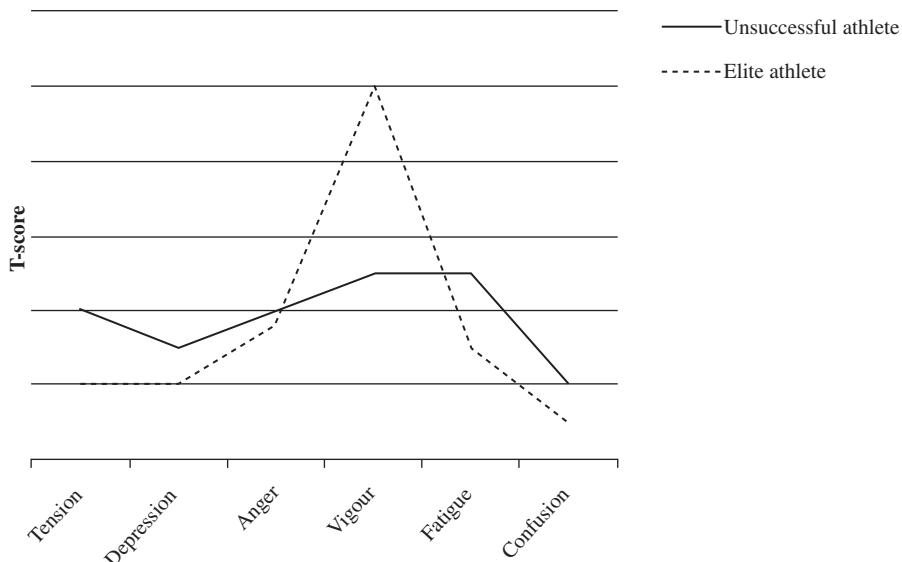


Figure 9.2 The Iceberg Profile.

The Iceberg profile is not without its critics who have challenged the validity of these findings and suggest that the observed differences between elite and non-elite athletes are not substantial (Leunes & Burger, 2000). Furthermore, Leunes and Burger (2000) argue the POMS is limited because it only considers negative mood states, essentially ignoring the role of positive mood states. Prapavessis (2000) argues the Iceberg Profile fails to consider individual mood states, and suggests an alternative approach is Hanin's Individual Zone of Optimal Function (IZOF; Hanin, 1997) model. This model takes into consideration both substandard and peak athletic performances to detect the athlete's individual content and intensity of emotional experiences in connection with their successes and failures (Hanin, 1997; Robazza, Pellizzari, Bertollo, & Hanin, 2008).

Current research has focused on coping strategies used by athletes to navigate the internal and external demands often encountered in the sport setting (Gaudreau & Blondin, 2002). The following are commonly used psychological instruments that attempt to measure such constructs.

## Coping skills

The transactional model of stress and coping developed by Lazarus and Folkman (1987) explained coping as a phenomenon that involves both cognitive and behavioral responses that individuals use in an attempt to manage internal and/or external stressors perceived to exceed their personal resources. It is safe to say that coping with stressful and anxiety-provoking situations is first on the list of an athlete's job description; and failure to do so will most likely result in subpar athletic performance. Omar-Fauzee, Daud, Abdullah, and Rashid (2009) observed that athletes at various at various levels of sport participation made use of a variety of effective coping strategies to help deal with the multiple stressors they encounter.

The Athletic Coping Skills Inventory (ACSI-28) developed by Smith, Schultz, Smoll, and Ptacek (1995) is a popular self-report measure used to assess the various coping techniques implemented by athletes. The ACSI-28 measures the following domains: Coping with Adversity, Peaking Under Pressure, Goal Setting and Mental Preparation, Concentration, Freedom from Worry, Confidence and Achievement Motivation, and Coachability. Spieler et al. (2007) studied coping differences between starters and non-starting collegiate varsity football players and concluded that when combined with other factors such as size of high school when growing up as well as age, athletes who were part of the starting line-up scored higher on the Coping with Adversity domain.

The Coping Inventory for Stressful Situation (CISS), a 48-item self-report measure developed by Endler and Parker (1999), uses a multidimensional approach to assess coping across three distinct orientations: Task-Focused, Emotion-Focused, and Avoidance-Focused. Philippe, Seiler, and Mengisen (2004) assessed male and

female elite athletes using the CISS and found female athletes scored higher in the use of Emotion and Avoidance Orientation compared to their male counterparts who scored higher in the use of Task Orientation as a coping approach. The authors attributed these observed coping behavior differences to the effect of social and gender norms, yet explained these observed differences were not substantial and only partly supported the “socialization model” (Philippe et al., 2004). Additionally, when comparing elite and non-elite athletes, they found elite athletes scored higher on the Avoidance Orientation Scale, and explained that these athletes may engage in the use of evasive techniques to discount or ignore particular stressors (i.e., money and expectations) in order to maintain top performance levels.

Another commonly used tool is the Coping Inventory for Competitive Sport (CICS; Gaudreau & Blondin, 2002). This instrument is a 39-item self-report measure that utilizes a 5-point Likert Scale to assess 10 different athletic coping strategies under three main scales that include: Task Orientation, Distraction Orientation, and Disengagement Orientation. Andrews and Chen (2014) utilized the CICS to study 478 runners of varying competencies to assess gender differences in coping styles during recovery from injury. They concluded that men used the Task-Oriented approach, particularly visualization, to cope with their respective injuries (“I can see myself getting better”). On the other hand, women more commonly adopted a Disengagement-Oriented approach to manage their recovery from injury (“I wish my injury would go away”). Nonetheless, they noted that the significant differences found in this study had low effect sizes.

## Resilience

Rutter (1985) defined resilience as a psychological phenomenon that utilizes protective factors that change someone’s ability to overcome stress and anxiety. In the sport setting, Fletcher and Sarkar (2012) argue resilience is a “prerequisite of sporting excellence.” They defined resilience as an athlete’s ability to accurately appraise when and how to use and optimize coping skills to confront stressful aspects associated with competition, the sports environment or daily life (Sarkar & Fletcher, 2014). Athletes who have the psychological ability to respond, react, and adapt to adversity are considered more likely to be successful (Galli & Gonzalez, 2015). Hence, athletes are not only expected to make use of these beneficial psychological skills but also to optimize them in order to withstand the pressures of their respective sports (Sarkar & Fletcher, 2014). Although Gonzalez and and collaborators (2016) and Sarkar and Fletcher (2013) note that measures of resilience in the sport setting continue to be a work in progress (Gonzalez, Moore, Newton, & Galli, 2016), there appears to be consensus that the 10-item Connor–Davidson Resilience Scale (CD-RISC) provides a reliable measure of resiliency in the sport setting and is the current best choice (Gonzalez et al., 2016).

## Mental toughness

Mental toughness commonly has been defined as a group of psychological characteristics that promote an individual's ability to reliably achieve and maintain performance to the upper limits of their capabilities (Mahoney, Ntoumanis, Mallett, & Gucciardi, 2014). Mahoney and collaborators (2014) described mental toughness as a psychological mechanism that permits individuals to *strive*, *survive*, and *thrive* under challenging circumstances. Specific to the sport setting, researchers define mental toughness as the ability to be consistently better than one's opponent by remaining *determined*, *focused*, *confident*, and *in control* when under pressure (Madrigal, Hamill, & Gill, 2013). Perhaps because so many definitions have been proposed, some authors argue that mental toughness continues to be one of the most misunderstood constructs within the sport community (Jones, Hanton, & Connaughton, 2002).

In recent years various mental toughness measures have been developed including:

1. The Psychological Performance Inventory (PPI) developed by Loehr in 1986, is considered to be the first mental toughness instrument. This self-report measure consists of 42 items that are aggregated into seven scales of six items each: self-confidence, attention control, negative energy, motivation, attitude control, positive energy, and imagery control (PPI; Loehr, 1986).
2. The Psychological Performance Inventory-A is the successor to the PPI and was developed by Golby, Sheard, and van Wersch (2007). The PPI-A includes a global measure of mental toughness and four subscales: determination, self-belief, positive control, and visualization (Golby et al., 2007).

However, it is important to note that researchers remain cautious about the utilization of these two particular measures due to methodological and psychometric flaws (Gucciardi, 2012).

Other frequently used mental toughness scales include the Mental Toughness Questionnaire (Clough, Earle, & Sewell, 2002), the Mental Toughness Scale (MTS; Madrigal et al., 2013), the Sports Mental Toughness Questionnaire (Sheard, Golby, & van Wersch, 2009), and the Mental, Emotional, and Bodily Toughness Inventory (Mack & Ragan, 2008). Madrigal and collaborators (2013) warn practitioners and researchers that all these measures are still in early developmental stages, have limited psychometric support, and lack evidence of construct validity.

## Team cohesiveness

When referring to team sports, it is important to assess an athlete's ability to adequately interact with and work alongside his/her teammates. Carron (1982) defined group cohesiveness as a dynamic process in which a group of individuals work together to achieve a common goal or objective. He noted that team cohesion involves the interaction of four distinct factors: environmental, personal, leadership, and team factors (Carron, 1982). The environmental factor relates to contractual responsibilities of a player within a sporting club or organization;

personal factors include an athlete's individual level of motivation, team satisfaction, and differences, while leadership factors subsume leader behavior and style, athlete–coach relationship, and the athlete–team relationship. Lastly, the team factor involves completion of group tasks, desire and ability to accomplish collective success in achieving a challenging task, group orientation, group productivity norm (high or low), and team stability (or how long has the group maintained together) (Caron, 1982). Commonly used team cohesion assessments can be found in Table 9.2.

### **Section key points**

1. There is no one measure that can predict elite athletic performance; however, test developers have been able to create distinct instruments that assess particular psychological constructs considered to be predictive of athletic success.
2. Psychological assessment instruments can provide adequate representation of an athlete's internal dynamics but the interaction of such constructs with other relevant environmental, physical, and cultural factors must be considered to obtain a complete picture of the athlete's capabilities.
3. Practitioners are encouraged to utilize instruments specific to sports participation that have been developed to assess the various constructs and behaviors often observed in this setting rather than making use of other, more common clinical psychology personality assessments.
4. Ethical considerations must be addressed prior, during, and after assessing athletes. The team as well as the individual athlete must understand any limitations of privilege and confidentiality of the information collected, as well as the interpretation.

## **Assessment in sports neuropsychology**

Sports neuropsychology is a relatively young subspecialty that grew out of the application of clinical neuropsychology to the sports domain. Unlike sports psychology where the assessment focus is on the enhancement of sports performance, the focus of sports neuropsychology is the assessment and management of sports-related brain injuries, particularly SRC. Inherent in the genesis and development of sports neuropsychology has been an unprecedented interest in SRC from research, clinical, and social-cultural arenas. Groups as diverse as the media, medical practitioners, governmental agencies, politicians, parents, professional and recreational sports leagues, and attorneys have all taken a keen interest in SRC. This interest has propelled sports neuropsychologists to be at the forefront of research in assessment, clinical intervention, and basic science (Echemendia & Bauer, 2015).

The Sports Neuropsychology Society defines sports neuropsychology as follows:

*A subspecialty of clinical neuropsychology that applies the science and understanding of brain–behavior relationships to the assessment and treatment of*

*sports-related brain injury. The practice of sports neuropsychology requires education, training, experience, and competence in the primary field of clinical neuropsychology, followed by a secondary specialization through experience and understanding of applying clinical neuropsychology to the unique demands of evaluating and treating brain injury in the sports domain (Sports Neuropsychology Society, 2014).*

## Concussion—the basics

A cerebral concussion is a brain injury often referred to as a mild traumatic brain injury (MTBI). Although there exist a wide range of definitions of MTBI in general and SRC specifically, according to the most recent international consensus statement published by the Concussion in Sport Group (CISG) (McCrory et al., 2017), a concussion is defined as follows:

*Sport related concussion is a traumatic brain injury induced by biomechanical forces. Several common features that may be utilized in clinically defining the nature of a concussive head injury include:*

- SRC may be caused either by a direct blow to the head, face, neck or elsewhere on the body with an impulsive force transmitted to the head.
- SRC typically results in the rapid onset of short-lived impairment of neurological function that resolves spontaneously. However, in some cases, signs and symptoms evolve over a number of minutes to hours.
- SRC may result in neuropathological changes, but the acute clinical signs and symptoms largely reflect a functional disturbance rather than a structural injury and, as such, no abnormality is seen on standard structural neuroimaging studies.
- SRC results in a range of clinical signs and symptoms that may or may not involve loss of consciousness. Resolution of the clinical and cognitive features typically follows a sequential course. However, in some cases symptoms may be prolonged.

*The clinical signs and symptoms cannot be explained by drug, alcohol, or medication use, other injuries (such as cervical injuries, peripheral vestibular dysfunction, etc.) or other comorbidities (e.g., psychological factors or coexisting medical conditions) [p. 2].*

SRC occur frequently at all levels of sport and across a broad age range, accounting for approximately 10% of all athletic injuries (Gessel, Fields, Collins, Dick, & Comstock, 2007). In 1998, the Centers for Disease Control estimated that approximately 300,000 sports-related MTBIs occurred each year (Thurman, Branche, & Sniezek, 1998). However, this estimate was subsequently increased to 1.6 to 3.8 million SRC per year (Langlois, Rutland-Brown, & Wald, 2006), which is widely believed to be an underestimate of the true prevalence of SRC.

SRCS are often associated with one or more symptoms, impaired balance, and/or cognitive deficits (Asken et al., 2016; Barr & McCrea, 2001; Collins et al., 1999; Delaney, Lacroix, Gagne, & Antoniou, 2001; Erlanger et al., 2001, 2003; Guskiewicz, Ross, & Macciocchi, Barth, Alves, Rimel, & Jane, 1996; Makdissi et al., 2001; Marshall, 2001; Matser, Kessels, Lezak, & Troost, 2001; McCrea et al., 2003; McCrea, 2001b; Peterson, Ferrara, Mrazik, Piland, & Elliott, 2003; Riemann & Guskiewicz, 2000). These problems can be measured using symptom scales (Lovell & Collins, 1998; Lovell et al., 2006; Lovell, Iverson, Collins, McKeag, & Maroon, 1999), balance testing (Guskiewicz et al., 2001; McCrea et al., 2003; Peterson et al., 2003; Riemann & Guskiewicz, 2000) and neuropsychological testing (Echemendia, Putukian, Mackin, Julian, & Shoss, 2001). All three assessment modalities can identify significant changes in the first few days following injury, generally with normalization over 1–3 weeks (Collins et al., 1999; Erlanger et al., 2001; Guskiewicz, 2001; Lovell et al., 2003; Macciocchi et al., 1996; McCrea et al., 2003). The presentation of symptoms and the rate of recovery can be variable, which reinforces the value of assessing all three areas as part of a comprehensive sport concussion program.

Neuropsychological assessment has been described as an important ‘cornerstone’ of concussion management (Aubry et al., 2002). Over the years, concussion management programs that use neuropsychological assessment to assist in clinical decision-making have been instituted in professional sports (Lovell, 2006, 2012; Lovell, Echemendia, & Burke, 2004), colleges (Collins, Lovell, & Echemendia, 2004; Echemendia et al., 2001; Schatz & Covassin, 2006), high schools (Pardini & Collins, 2006) and school-age children (Brooks, 2006). In addition to neuropsychological measures, rapid cognitive screening tests have been developed (e.g., Standardized Assessment of Concussion, SAC) that are sensitive to the immediate effects of concussion (McCrea, 2001a; McCrea, Kelly, Randolph, Cisler, & Berger, 2002). The SAC, embedded in the Sport Concussion Assessment Tool (SCAT), can be used on the sideline or later on the day of injury to identify cognitive deficits associated with concussion (McCrory et al., 2009; McCrory et al., 2013). Although useful on the day of injury and in the few days that follow, brief cognitive screening tests such as the SAC or SCAT are not substitutes for more comprehensive neuropsychological assessment (McCrory et al., 2009, 2013). Numerous studies have found that both traditional (paper and pencil) and computerized cognitive tests are sensitive to the acute effects of concussion (Collins et al., 1999; Guskiewicz et al., 2001; Macciocchi et al., 1996; Makdissi et al., 2001; Matser et al., 2001; McCrea et al., 2003; Lovell et al., 2003). Consequently, many position and consensus statements have recommended neuropsychological assessment as an essential component of concussion management programs (Aubry et al., 2002; McCrory et al., 2005, 2009, 2013, 2017; Moser et al., 2007).

In this section, the various assessment tools that are commonly implemented in sports concussion management will be reviewed. Sideline evaluations and office evaluations will be highlighted, and the strengths and limitations of these approaches will be discussed.

## Assessment of concussion

The assessment of SRC typically begins with an acute evaluation on the field followed by a sideline or locker room evaluation, a formal post-acute neurocognitive assessment, graded progression of physical exertion, and finally unrestricted return to play. Each step is designed to answer a different set of clinical questions for which different instruments and techniques need to be used.

### **Sideline evaluations**

A sideline or on-field clinical examination of players is a critical first step. The primary goal of the acute “on-field” assessment is to identify any life-threatening conditions (e.g., developing intracranial bleeding) and to assess for the possibility of spinal cord injury. If an athlete’s symptoms are deteriorating, especially if there is deterioration to a stuporous, semicomatoso, or comatoso state of consciousness, the situation must be treated as a medical emergency, and emergency transport is required (Guskiewicz, Echemendia, & Cantu, 2009). If the athlete is deemed medically stable but a concussion is suspected, then a comprehensive sideline assessment should be conducted. Such assessment includes a thorough history, observation of signs and symptoms, player report of symptoms, a careful assessment of the player’s recall of the events prior to and following the injury, and assessment of the cognitive and physical areas that are frequently affected by concussion, including tests of learning and memory, concentration, motor coordination, and cranial nerve functioning.

Over the years there has been momentum towards using standardized, empirically validated brief screening tools to evaluate post-concussion signs and symptoms, cognitive functioning, postural stability, and more recently ocular-motor functioning on the sidelines immediately after concussion (Barr & McCrea, 2001; Galetta et al., 2015; McCrea et al., 2002). The SAC was introduced in 1997 and included a brief assessment of neurocognitive and neurological functions (McCrea et al., 2002, 1998; McCrea, Kluge, Ackley, & Randolph, 1997) designed to help standardize the sideline assessment. Following the 2004 CISG meeting in Prague, the SCAT (SCAT; McCrory et al., 2005) was introduced. The SCAT was designed as a measure that could be used to provide a rapid multifaceted and objective SAC. This new tool incorporated the SAC in addition to a graded symptom checklist (the Post-Concussion Symptom Scale), orientation questions specific to sport (modified Maddocks’ questions), an assessment of loss of consciousness and amnesia, and return to play guidelines (McCrory et al., 2005). The SCAT was widely implemented in sports concussion assessment at that time, and was later revised to the SCAT2 in 2008 (SCAT2; McCrory, 2009) based on a review of the available empirical literature. The revision included the addition of the Glasgow Coma Scale, alternate word lists for the SAC, and an assessment of balance (i.e., the modified Balance Error Scoring System—m-BESS). The revisions of the SCAT2 allowed for the measure to be used not only for sideline assessment but also for tracking

recovery over time (i.e., serial assessment). In addition to the SCAT2, which was designed for use by medical practitioners, a new tool was developed for use by non-medically trained individuals: the PocketSCAT2.

After another systematic review of the literature (Guskiewicz et al., 2013), the SCAT2 was revised in 2013 (McCrory et al., 2013), which produced the SCAT3 and a new tool for children under the age of 13, the Child SCAT3. The SCAT3 improved on the SCAT2 by including physical or objective signs of concussion, allowing for the addition of a more sensitive foam condition on the BESS, and additional changes to the “Concussion injury advice” section (Guskiewicz et al., 2013). The components of the SCAT3 include: Indications for emergency management, potential signs of concussion, Glasgow Coma Scale, Maddocks questions (sport-specific orientation questions), medical background questions, symptom evaluation, cognitive assessment, neck balance, and coordination examination, SAC delayed recall, considerations for management, and concussion advice. More recently, the SCAT5 was introduced for adolescents and adults (Echemendia et al., 2017, 2017b) and the Child SCAT5 for children 12 and younger (Davis et al., 2017). Both the SCAT5 and the Child SCAT5 made significant improvements over the SCAT3, including adding the option of using a 10 word list learning subtest as opposed to the five word list that was used in the SCAT3, which was shown to have significant ceiling effects (Echemendia et al., 2017). In addition to the SCAT5 and the Child SCAT5, which are designed for use by medically trained individuals, the Concussion Recognition Tool 5 (CRT5; Echemendia et al., 2017a) was introduced for use by lay personnel and emphasizes a “Recognition and Recovery” approach rather than assessment of the injury.

The SCAT and its subsequent iterations have been adopted in various forms by a wide range of professional sports organizations (e.g., NHL, NFL, MLB) and have been used extensively at the college and high school levels of sport participation. Despite the widespread use, it is important to note that brief measures such as the SCAT—while useful for obtaining an initial assessment of cognitive functioning during the acute phase of the injury—are not a substitute for formal neuropsychological assessment, which is usually conducted in the subacute phase of recovery (Aubry et al., 2002; McCrea et al., 2009). Furthermore, as these sideline assessments gain popularity, it is necessary to evaluate and understand the clinical utility of the measures. Since the first iteration of the SCAT, numerous studies have been conducted examining the psychometric properties of the measure, as well as its sensitivity and specificity.

A recent study by Chin and colleagues (Chin, Nelson, Barr, McCrory, & McCrea, 2016) evaluated the reliability and validity of the three main subtests of the SCAT3 (i.e., symptom checklist, cognitive assessment, and balance/postural stability assessment) in high school and college athletes. Results showed that among the three subtests, the symptom checklist demonstrated the greatest sensitivity to the effects of concussion. The authors concluded that symptom burden remains a key component of concussion assessment, but noted that it would be worthwhile for future research to determine circumstances under which cognitive assessment and balance testing might be more sensitive to the effects of concussion.

Examining individual differences on SCAT performance has been another area of interest. Chin et al. (2016) reported on individual differences observed at baseline on the SCAT3. With respect to sex differences, the authors found that female athletes reported greater symptoms than males, but also demonstrated stronger baseline cognitive performance when compared to males. This finding is consistent with several past studies examining the SCAT and SCAT2, with females outperforming males on cognitive aspects but endorsing a greater symptom burden (Jinguji et al., 2012; Shehata et al., 2009; Valovich McLeod, Barr, McCrea, & Guskiewicz, 2006). With respect to age, findings have been mixed. There is some evidence to support the notion that younger high school athletes endorse fewer symptoms and outperform older high school athletes on the cognitive portion of the SCAT2 (Jinguji et al., 2012). However, others have shown that total SCAT2 scores increase with age (Valovich McLeod et al., 2012; Snyder & Bauer, 2014). Among children, age has been a consistent finding with older children outperforming younger children (Davis et al., 2017).

As for ADHD/LD status, athletes with these conditions endorsed greater symptoms and exhibited a worse cognitive performance compared to athletes without ADHD or LD (Chin et al., 2016). Finally, although the influence of concussion history on SCAT performance has been well documented, disparate results have been reported. Chin et al. (2016) found that concussion history did not have an effect on SCAT3 baseline performance. Similar results have been reported for the SCAT2 (Hanninen et al., 2016; Zimmer, Marcinak, Hibyan, & Webbe, 2015). Yet, other authors have found that athletes with a history of concussion performed more poorly on the SCAT2 than athletes without a history of concussion (Valovich McLeod et al., 2012).

### **Off-field evaluations**

The introduction of neuropsychological assessment in sports can be traced to the early work of Jeff Barth and his colleagues who introduced the baseline-post-injury evaluation model that continues to be used today. Barth used a limited battery of traditional “paper and pencil” neuropsychological tests to assess college football players before and after sustaining a SRC (Barth et al., 1989). Computerized batteries were developed in the 1990s to provide an alternative to traditional tests and are now used almost exclusively in many sports settings. Traditional tests have been studied in combination with computerized batteries to assess construct validity (Maeirlender et al., 2010; Maruff et al., 2009). The combined use of traditional and computerized neuropsychological tests in applied settings has been referred to as a ‘hybrid’ neuropsychological testing approach (Echemendia, 2012).

There is no scientific evidence that traditional tests, computerized tests or a hybrid approach is superior; each approach has its strengths and limitations. Traditional tests are reasonably reliable, valid and sensitive to the effects of sports concussion (Macciocchi et al., 1996; Randolph, McCrea, & Barr, 2005; Echemendia et al., 2001). They can be selected to fit the specific needs of the athlete and

domains of neuropsychological importance. These tests have a much longer history of being applied in clinical settings, with some having large normative databases. However, traditional tests require face-to-face examination, which may introduce variance in test administration and scoring. These tests are also more labor-intensive, especially in sports settings where group testing is more efficient given the large numbers of athletes involved.

In contrast, computerized tests such as the Immediate Post-concussion Assessment and Cognitive Testing (ImPACT), Axon Sports, Automated Neuropsychological Assessment Metrics (ANAM), Headminder, and C3 can be rapidly administered to individuals or groups. Computerized batteries are portable and efficient for the collection, synthesis, and storage of large amounts of data. Clinical end-user reports are often immediately available.

Although computerized tests offer certain efficiencies, they also have limitations. First, they are brief and rely on a limited sample of cognitive functioning. Second, although a number of studies have shown that computerized batteries appear to have adequate psychometric properties (Bleiberg et al., 2004; Broglio, Ferrara, Macciocchi, Baumgartner, & Elliott, 2007; Collie, Darby, & Maruff, 2001; Collie et al., 2003; Collins et al., 2003; Iverson, Lovell, & Collins, 2005; Schatz, Pardini, Lovell, Collins, & Podell, 2006), other studies have raised questions about the reliability, form equivalence and validity of these instruments (Broglio et al., 2007; Mayers & Redick, 2012; Resch et al., 2013). For instance, test-retest reliabilities have been reported to be quite low by some investigators, which increases reliable change (RC) metrics and limits the value of baseline examinations (Bruce et al., 2016; Bruce, Echemendia, Meeuwisse, Comper, & Sisco, 2013; Echemendia et al., 2016; Randolph et al., 2005; Mayers & Redick, 2012; Broglio et al., 2007; Resch et al., 2013). Third, alternate forms may not be equivalent (Resch et al., 2013). Fourth, while one of the major purported advantages of computerized testing is the application to group testing, recent research (Moser, Schatz, Neidzwski, & Ott, 2011) indicates that group versus individual test administration may yield different results. Based on the combined experience of the present authors, under most circumstances large group baseline testing should be avoided because of an inability to adequately control the testing environment. Fifth, although computers have provided a technological breakthrough in neurocognitive testing, the technology itself is subject to error. Significant variability exists in the accuracy of personal computers' measurement of response time (Schatz, 2014), although they claim to do so with millisecond accuracy. Variability has also been found in the use of computer mice or keyboards, and monitor refresh rates (Cernich, Brennana, Barker, & Bleiberg, 2007). Sixth, computerized batteries have been marketed to clinicians from diverse disciplines, some of whom may have no education or training in cognitive assessment (Bauer et al., 2012; Echemendia, Herring, & Bailes, 2009). Finally, computerized batteries may be seen as a 'black box' approach to neuropsychological assessment, an approach that is partially encouraged by the immediate availability of a clinical report that contains simplified coding for whether a finding is reliable or significant. This may lead some clinicians astray by inaccurately providing a singular

or ‘cookbook’ approach to assessment that is complicated by many of the factors discussed above.

Some authors have advocated the hybrid approach (Echemendia et al., 2013), which combines the benefits of both computer based batteries and traditional tests. Typically, the hybrid approach uses a computerized battery at baseline and a combination of computerized and traditional tests following injury. Although the hybrid method is intuitively appealing, only a few studies to date have examined the clinical utility of this approach (Maerlender et al., 2010; Maruff et al., 2009).

Even though there has been considerable excitement regarding the use of neuropsychological tests in the evaluation of SRC, there are also those who have been critical of the field (Kirkwood, Randolph, & Yeates, 2009; Randolph & Kirkwood, 2009; Randolph et al., 2005). The criticisms set forth areas that are in need of further research and questions that need to be answered more completely. One area of concern is the widespread use of baseline testing. Although widely adopted, there are no studies that have established whether baseline testing adds greater precision to the detection of post-injury cognitive deficits when compared to post-injury evaluations alone (Echemendia et al., 2012). Although the use of baseline testing appears attractive and even logical, it does not come without costs because it introduces significant complexity into the interpretation of post-injury test data. The critical *theoretical* advantage of baseline testing is the reduction of within-subject variability. Unfortunately, the extent to which baseline and post-injury assessments do not employ similar characteristics of environment and examiner at each time point quickly undermines the rationale for using baseline methods. Additionally, not only is there error surrounding the post-injury tests, there is also error around the baseline tests as well as the error associated with comparing tests at two different time intervals, particularly in light of poor temporal stability found among some of these tests (Barr, 2003; Broglio et al., 2007; Mayers & Redick, 2012). The widespread use of computerized testing has created the perception that little or no formal training is needed to administer and interpret these tests, which leads to the basic question, “Who should administer and interpret neuropsychological tests?” Many programs, perhaps most programs, have adopted a model where tests are administered and interpreted without consultation of a qualified neuropsychologist. The CISG (McCrory et al., 2017) concluded, “Neuropsychological (NP) assessment has been previously described by the CISG as a ‘cornerstone’ of SRC management. Neuropsychologists are uniquely qualified to interpret NP tests and can play an important role within the context of a multifaceted-multimodal and multidisciplinary approach to managing SRC” (p. 4). Echemendia et al. (2009) examined this question at length and similarly concluded, “The interpretation of neuropsychological tests requires comprehensive knowledge of the tests, their characteristics given a specific population (e.g., team, sport), the athlete and his or her specific situation, psychological variables and many others. For these reasons we conclude that neuropsychological tests may be administered under the guidance of a neuropsychologist but that the interpretation of neuropsychological test data is best managed by a clinical neuropsychologist.”

### **Section key points**

1. SRC are complex injuries that require multimodal assessment and management by a multidisciplinary team using a biopsychosocial framework.
2. Neuropsychological assessment is central to the evaluation of SRC and has evolved to include both traditional test batteries and computerized testing platforms, each with their respective strengths and benefits.
3. Neuropsychological assessment in the sports context has developed quickly over the last decade with significant challenges still ahead including further development, refinement, and psychometric enhancement of current test batteries. Of critical importance is the addition of multi-language versions of tests, particularly at the level of professional sports ([Echemendia, Bruce, & Glusman, 2018](#)).

## **Summary and conclusions**

This chapter highlighted psychological and neuropsychological approaches utilized in the context of sports assessment. The first section reviewed the various disciplines of sports psychology, the unique aspects of the sport environment, psychological approaches to assessment in sport psychology, and commonly measured constructs and behaviors such as personality, resilience, and mental toughness. Over the years, several unique measures and tools have been developed that allow for a more precise assessment of predictors of athletic performance. Consequently, sports psychologists are now better suited to be able to address challenging questions pertaining to psychological mechanisms that influence athletes' achievements and abilities. The second section of the chapter highlighted assessment approaches in sports neuropsychology, with an emphasis on SRC. The clinical signs and symptoms of concussion were discussed, followed by a detailed explanation of the sideline and off-field evaluations that are commonly conducted following SRC. The advantages and disadvantages associated with the various assessment modalities were described, and current evidence regarding the use of the baseline testing approach was examined. Although more research is clearly necessary to determine the best method for the assessment and management of SRC, it is clear that the field of sports neuropsychology has made significant advances over the past several years.

## **References**

- Andrews, P., & Chen, M. A. (2014). Gender differences in mental toughness and coping with injury in runners. *Journal of Athletic Enhancement*, 3, 6.
- American Psychological Association Division 47. (2009). What is exercise and sport psychology? Retrieved from <http://www.apa47.org/pracExSpPsych.php>.
- American Psychological Association. (2010). Clinical psychology. Retrieved from <http://www.apa.org/ed/graduate/specialize/neuro.aspx>.
- Anshel, M. H. (1990). Toward validation of a model for coping with acute stress in sport. *International Journal of Sport Psychology*, 21, 58–83.

- Asken, B. M., McCrea, M. A., Clugston, J. R., Snyder, A. R., Houck, Z. M., & Bauer, R. M. (2016). "Playing through it": Delayed reporting and removal from athletic activity after concussion predicts prolonged recovery. *Journal of Athletic Training*, 51(4), 329–335. Available from <https://doi.org/10.4085/1062-6050-51.5.02>.
- Aubry, M., Cantu, R., Dvorak, J., Graf-Baumann, T., Johnston, K., Kelly, J. P., ... Schamasch, P. (2002). Concussion in sport group. Summary and agreement statement of the first international conference on concussion in sport, Vienna 2001. *British Journal of Sports Medicine*, 36, 6–10.
- Barr, W. B. (2003). Neuropsychological testing of high school athletes. Preliminary norms and test-retest indices. *Archives of Clinical Neuropsychology*, 18, 91–101.
- Barr, W. B., & McCrea, M. (2001). Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion. *Journal of the International Neuropsychological Society : JINS*, 7(6), 693–702.
- Barth, J. T., Alves, W. M., Ryan, T. V., Macciocchi, S. N., Rimel, R. E., & Jane, J. A. (1989). Mild head injury in sports: Neuropsychological sequelae and recovery of function. In H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Mild head injury* (pp. 257–275). New York: Oxford University Press.
- Bauer, R. M., Iverson, G., Cernich, A. N., Binder, L. M., Ruff, R., & Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Archives of Clinical Neuropsychology*, 27, 362–373.
- Bleiberg, J., Cernich, A. N., Cameron, K., Sun, W., Peck, K., Ecklund, P. J., & Warden, D. L. (2004). Duration of cognitive impairment after sports concussion. *Neurosurgery*, 54, 1073–1078.
- Boring, E. G. (1929). *History of experimental psychology*. New York: Appleton-Century.
- Brinthaupt, T. M., & Anshel, M. H. (2015). Practical guidelines for using inventories in sport psychology. *The Sport and Exercise Scientist*, 45, 12–13.
- Broca, P. P. (1861). Loss of speech, chronic softening and partial destruction of the anterior left lobe of the brain. *Bulletin de la Société Anthropolistique*, 2, 235–238.
- Broglio, S. P., Ferrara, M. S., Macciocchi, S. N., Baumgartner, T. A., & Elliott, R. (2007). Test-retest reliability of computerized concussion assessment programs. *Journal of Athletic Training*, 42(4), 509–514.
- Brooks, J. (2006). Concussion management programs for school-age children. In R. Echemendia (Ed.), *Sports neuropsychology: Assessment and management of traumatic brain injury* (pp. 131–141). New York: Guilford Press.
- Bruce, J., Echemendia, R., Tangeman, L., Meeuwisse, W., Comper, P., Hutchison, M. G., & Aubry, M. (2016). Two baselines are better than one: Improving the reliability of computerized testing in sports neuropsychology. *Applied Neuropsychology*. Available from <https://doi.org/10.1080/23279095.2015.1064002>.
- Bruce, J., Echemendia, R. J., Meeuwisse, W., Comper, P., & Sisco, A. (2013). One year test-retest reliability of ImPACT in professional ice hockey players. *The Clinical Neuropsychologist*, 28, 14–24.
- Bush, S. S., & Iverson, G. L. (2011). Ethical issues and practical considerations. In F. M. Webbe (Ed.), *Handbook of sport neuropsychology* (pp. 35–52). New York: Springer Publishing Company.
- Carron, A. V. (1982). Cohesiveness in sport groups: Interpretations and considerations. *Journal of Sport Psychology*, 4(2), 123–138.
- Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (1993). *16PF fifth edition questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

- Cernich, A. N., Brennana, D. M., Barker, L. M., & Bleiberg, J. (2007). Sources of error in computerized neuropsychological assessment. *Archives of Clinical Neuropsychology*, 22, s39–48.
- Chin, E. Y., Nelson, L. D., Barr, W. B., McCrory, P., & McCrea, M. A. (2016). Reliability and validity of the sport concussion assessment tool-3 (SCAT3) in high school and collegiate athletes. *American Journal of Sports Medicine*, 44(9), 2276–2285. Available from <https://doi.org/10.1177/0363546516648141>.
- Clough, P., Earle, K., & Sewell, D. (2002). Mental toughness: The concept and its measurement. In I. Cockerill (Ed.), *Solutions in sport psychology* (pp. 32–45). London, England: Thomson.
- Collie, A., Darby, D., & Maruff, P. (2001). Computerised cognitive assessment of athletes with sports related head injury. *British Journal of Sports Medicine*, 35(5), 297–302.
- Collie, A., Maruff, P., Makdissi, M., McCrory, P., McStephen, M., & Darby, D. (2003). CogSport: Reliability and correlation with conventional cognitive tests used in postconcussion medical evaluations. *Clinical Journal of Sport Medicine*, 13(1), 28–32.
- Collins, M., Lovell, M., & Echemendia, R. (2004). Models of neuropsychological assessment —collegiate and high school sports. In M. Lovell, R. Echemendia, J. Barth, & M. Collins (Eds.), *Traumatic brain injury in sports: An international neuropsychological perspective* (pp. 479–499). Lisse: Zwets-Zeitlinger.
- Collins, M. W., Field, M., Lovell, M. R., Iverson, G., Johnston, K. M., Maroon, J., & Fu, F. H. (2003). Relationship between postconcussion headache and neuropsychological test performance in high school athletes. *American Journal of Sports Medicine*, 31(2), 168–173.
- Collins, M. W., Grindel, S. H., Lovell, M. R., Dede, D. E., Moser, D. J., Phalin, B. R., ... McKeag, D. B. (1999). Relationship between concussion and neuropsychological performance in college football players. *JAMA: The Journal of the American Medical Association*, 282(10), 964–970.
- Connor, K. M., & Davidson, J. R. T. (2003). Development of a new resilience scale: The Connor—Davidson Resilience Scale (CD-RISC). *Depression and Anxiety*, 18(2), 76–82.
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and the NEO five-factor inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.
- Davis, G. A., Purcell, L., Schneider, K., Yeates, K. O., Gioia, G., Anderson, V., ... Meeuwisse, W. (2017). The child sport concussion assessment tool 5th edition (Child-SCAT5). *British Journal of Sports Medicine*, 51(11), 859–861.
- Delaney, J. S., Lacroix, V. J., Gagne, C., & Antoniou, J. (2001). Concussions among university football and soccer players: a pilot study. *Clinical Journal of Sport Medicine: Official Journal of the Canadian Academy of Sport Medicine*, 11(4), 234–240.
- Echemendia, R. (2012). Cerebral concussion in sport: An overview. *Journal of Clinical Sport Psychology*, 6, 207–230.
- Echemendia, R., Bruce, J., Bailey, C. M., Sanders, F., Arnett, P., & Vargas, G. (2012). The utility of post-concussion neuropsychological data in identifying cognitive change in the absence of baseline data. *The Clinical Neuropsychologist*, 26(7), 1077–1091.
- Echemendia, R., Bruce, J., Meeuwisse, W., Comper, P., Aubry, M., & Hutchison, M. (2016). Long-term reliability of ImPACT in professional ice hockey. *The Clinical Neuropsychologist*. Available from <https://doi.org/10.1080/13854046.2016.1158320>.
- Echemendia, R. J., & Bauer, R. M. (2015). Professional ethics in sports neuropsychology. *Psychological Injury and Law*, 8(4), 289–299.
- Echemendia, R. J., Broglio, S. P., Davis, G. A., Guskiewicz, K., Hayden, K. A., Leddy, J., ... McCrory, P. (2017). What tests and measures should be added to the SCAT3 and

- related tests to improve their reliability, sensitivity and/or specificity in sideline diagnosis? A systematic review. *British Journal of Sports Medicine*.
- Echemendia, R.J., Bruce, J.M. & Glusman (2018). Professional sports neuropsychology. In P. Arnett (Ed.), *Neuropsychological perspectives on sports-related concussion*. Washington, DC: American Psychological Association.
- Echemendia, R. J., Herring, S., & Bailes, J. (2009). Who should administer and interpret neuropsychological tests? *British Journal of Sports Medicine*, 43, 32–35.
- Echemendia, R. J., Iverson, G., McCrea, M., Macciocchi, S. N., Gioia, G., & Putukian, M. (2013). Advances in neuropsychological assessment. *British Journal of Sports Medicine*, 47, 294–298.
- Echemendia, R. J., Meeuwisse, W., McCrory, P., Davis, G., Putukian, M., Leddy, J., ... Herring, S. (2017a). The concussion recognition tool 5th edition (CRT5). *British Journal of Sports Medicine*, 51, 870–871.
- Echemendia, R. J., Meeuwisse, W., McCrory, P., Davis, G., Putukian, M., Leddy, J., ... Herring, S. (2017b). The Sport Concussion Assessment Tool 5th Edition (SCAT5). *British Journal of Sports Medicine*, 51, 848–850.
- Echemendia, R. J., Putukian, M., Mackin, S., Julian, L., & Shoss, N. (2001). Neuropsychological test performance prior to and following sports-related mild traumatic brain injury. *Clinical Journal of Sport Medicine*, 11, 23–31.
- Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: The Dryden Press.
- Endler, N. S., & Parker, J. D. A. (1999). *Coping inventory for stressful situations (CISS): Manual* (2nd ed.). Toronto, ON: Multi-Health System.
- Erlanger, D., Kaushik, T., Cantu, R., Barth, J. T., Broshek, D. K., Freeman, J. R., & Webbe, F. M. (2003). Symptom-based assessment of the severity of a concussion. *Journal of neurosurgery*, 98(3), 477–484. Available from <https://doi.org/10.3171/jns.2003.98.3.0477>.
- Erlanger, D., Saliba, E., Barth, J., Almquist, J., Webright, W., & Freeman, J. (2001). Monitoring resolution of postconcussion symptoms in athletes: Preliminary results of a web-based neuropsychological test protocol. *Journal of Athletic Training*, 36(3), 280–287.
- Fletcher, D., & Sarkar, M. (2012). A grounded theory of psychological resilience in Olympic champions. *Psychology of Sport and Exercise*, 13(5), 669–678.
- Galetta, K. M., Morganroth, J., Moehringer, N., Mueller, B., Hasanaj, L., Webb, N., ... Balcer, L. J. (2015). Adding vision to concussion testing: A prospective study of sideline testing in youth and collegiate athletes. *Journal of Neuro-Ophthalmology*, 35(3), 235–241. Available from <https://doi.org/10.1097/WNO.0000000000000226>.
- Galli, N., & Gonzalez, S. P. (2015). Psychological resilience in sport: A review of the literature and implications for research and practice. *International Journal of Sport and Exercise Psychology*, 13(3), 243–257.
- Gaudreau, P., & Blondin, J. P. (2002). Development of a questionnaire for the assessment of coping strategies employed by athletes in competitive sport settings. *Psychology of Sport and Exercise*, 3(1), 1–34.
- Gessel, L. M., Fields, S. K., Collins, C. L., Dick, R. W., & Comstock, R. D. (2007). Concussions among United States high school and collegiate athletes. *Journal of Athletic Training (National Athletic Trainers' Association)*, 42(4), 495–503.
- Golby, J., Sheard, M., & van Wersch, A. (2007). Evaluating the factor structure of the psychological performance inventory. *Perceptual and Motor Skills*, 105, 309–325.
- Gonzalez, S. P., Moore, E. W. G., Newton, M., & Galli, N. A. (2016). Validity and reliability of the Connor–Davidson Resilience Scale (CD-RISC) in competitive sport. *Psychology of Sport and Exercise*, 23, 31–39.

- Gucciardi, D. F. (2012). Measuring mental toughness in sport: A psychometric examination of the psychological performance inventory—A and its predecessor. *Journal of Personality Assessment*, 94(4), 393–403.
- Gucciardi, D., & Gordon, S. (2011). *Mental toughness in sport: Developments in theory and research*. New York: Routledge.
- Guskiewicz, K., Echemendia, R., & Cantu, R. (2009). Assessment and return to play following sports-related concussion. In W. B. Kibler (Ed.), *Orthopaedic knowledge update: Sports medicine* (pp. 285–294). Rosemont, IL: American Academy of Orthopaedic Surgeons.
- Guskiewicz, K. M. (2001). Postural stability assessment following concussion: One piece of the puzzle. *Clinical Journal of Sport Medicine*, 11(3), 182–189.
- Guskiewicz, K. M., Register-Mihalik, J., McCrory, P., McCrea, M., Johnston, K., Makdissi, M., ... Meeuwisse, W. (2013). Evidence-based approach to revising the SCAT2: Introducing the SCAT3. *British Journal of Sports Medicine*, 47(5), 289–293. Available from <https://doi.org/10.1136/bjsports-2013-092225>.
- Guskiewicz, K. M., Ross, S. E., & Marshall, S. W. (2001). Postural stability and neuropsychological deficits after concussion in collegiate athletes. *Journal of Athletic Training*, 36(3), 263–273.
- Gutmann, M. C., Pollock, M. L., Foster, C., & Schmidt, D. (1984). Training stress in Olympic speed skaters: A psychological perspective. *The Physician and Sports Medicine*, 12(12), 45–57.
- Hanin, Y. L. (1997). Emotions and athletic performance: Individual zones of optimal functioning model. *European Yearbook of Sport Psychology*, 1, 29–72.
- Hanninen, T., Tuominen, M., Parkkari, J., Vartiainen, M., Ohman, J., Iverson, G. L., & Luoto, T. M. (2016). Sport concussion assessment tool—3rd edition—normative reference values for professional ice hockey players. *Journal of Science and Medicine in Sport*, 19(8), 636–641.
- Heuchert, J. P., & McNair, D. M. (2012). *The profile of mood states 2nd edition (POMS 2)*. North Tonawanda, NY: Multi-Health Systems.
- Iverson, G. L., Lovell, M., & Collins, M. (2005). Validity of ImPACT for measuring processing speed following sports-related concussion. *Journal of Clinical and Experimental Neuropsychology*, 27(6), 683–689.
- Jinguiji, T. M., Bompadre, V., Harmon, K. G., Satchell, E. K., Gilbert, K., Wild, J., & Eary, J. F. (2012). Sport Concussion Assessment Tool-2: Baseline values for high school athletes. *British Journal of Sports Medicine*, 46(5), 365–370. Available from <https://doi.org/10.1136/bjsports-2011-090526>.
- Jones, G., Hanton, S., & Connaughton, D. (2002). What is this thing called mental toughness? An investigation of elite performers. *Journal of Applied Sport Psychology*, 14, 205–218.
- Kirkwood, M. W., Randolph, C., & Yeates, K. O. (2009). Returning pediatric athletes to play after concussion: The evidence (or lack thereof) behind baseline neuropsychological testing. *Acta Paediatrica*, 98, 1409–1411.
- Langlois, J. A., Rutland-Brown, L. V., & Wald, M. M. (2006). The epidemiology and impact of traumatic brain injury. *Journal of Head Trauma Rehabilitation*, 21, 375–378.
- Lazarus, R. S., & Folkman, S. (1987). Transactional theory and research on emotions and coping. *European Journal of Personality*, 1(3), 141–169.
- Leunes, A., & Burger, J. (2000). Profile of mood states research in sport and exercise psychology: Past, present, and future. *Journal of Applied Sport Psychology*, 12(1), 5–15.
- Lezak, M. D. (1983). *Neuropsychological assessment* (2nd ed). New York: Oxford University Press.

- Li, F. (1999). The exercise motivation scale: Its multifaceted structure and construct validity. *Journal of Applied Sport Psychology, 11*(1), 97–115.
- Loehr, J. E. (1986). *Mental toughness training for sports: Achieving athletic excellence*. Lexington, MA: Stephen Greene Press.
- Lovell, M., Echemendia, R., & Burke, C. (2004). Professional ice hockey. In M. Lovell, R. Echemendia, J. Barth, & M. Collins (Eds.), *Mild traumatic brain injury in sports: An international perspective* (pp. 221–231). Amsterdam: Zwets-Zeitlinger.
- Lovell, M. R. (2006). *Neuropsychological assessment of the professional athlete. Sports neuropsychology: Assessment and management of traumatic brain injury* (pp. 176–190). New York, NY, US: Guilford Press.
- Lovell, M. R. (2012). *Assessment of mild traumatic brain injury in the professional athlete. Neuropsychological assessment of work-related injuries* (pp. 68–79). New York, NY, US: Guilford Press.
- Lovell, M. R., & Collins, M. W. (1998). Neuropsychological assessment of the college football player. *Journal of Head Trauma Rehabilitation, 13*(2), 9–26.
- Lovell, M. R., Collins, M. W., Iverson, G. L., Field, M., Maroon, J. C., Cantu, R., ... Fu, F. H. (2003). Recovery from mild concussion in high school athletes. *Journal of Neurosurgery, 98*(2), 296–301. Available from <https://doi.org/10.3171/jns.2003.98.2.0296>.
- Lovell, M. R., Iverson, G. L., Collins, M. W., McKeag, D., & Maroon, J. C. (1999). Does loss of consciousness predict neuropsychological decrements after concussion? *Clinical Journal of Sport Medicine, 9*(4), 193–198.
- Lovell, M. R., Iverson, G. L., Collins, M. W., Podell, K., Johnston, K. M., Pardini, D., ... Maroon, J. C. (2006). Measurement of symptoms following sports-related concussion: Reliability and normative data for the post-concussion scale. *Applied Neuropsychology, 13*(3), 166–174. Available from [https://doi.org/10.1207/s15324826an1303\\_4](https://doi.org/10.1207/s15324826an1303_4).
- Macciocchi, S. N., Barth, J. T., Alves, W., Rimel, R. W., & Jane, J. A. (1996). Neuropsychological functioning and recovery after mild head injury in collegiate athletes. *Neurosurgery, 39*(3), 510–514.
- Mack, M. G., & Ragan, B. G. (2008). Development of the mental, emotional, and bodily toughness inventory in collegiate athletes and nonathletes. *Journal of Athletic Training, 43*(2), 125.
- Madrigal, L., Hamill, S., & Gill, D. L. (2013). Mind over matter: The development of the Mental Toughness Scale (MTS). *The Sport Psychologist, 27*, 62–77.
- Maerlender, A., Flashman, L., Kessler, A., Kumbhani, S., Greenwald, R. M., Tosteson, T., & McAllister, T. (2010). Examination of the construct validity of ImPACT™ computerized test, traditional, and experimental neuropsychological measures. *Clinical Neuropsychologist, 24*(8), 1309–1325. Available from <https://doi.org/10.1080/13854046.2010.516072>.
- Mahoney, J., Ntoumanis, N., Mallett, C., & Gucciardi, D. (2014). The motivational antecedents of the development of mental toughness: A self-determination theory perspective. *International Review of Sport and Exercise Psychology, 7*(1), 184–197.
- Mainwaring, L. (2011). Short-term and extended emotional correlates of concussion. In F. M. Webbe (Ed.), *Handbook of sport neuropsychology* (pp. 251–273). New York: Springer Publishing Company.
- Makdissi, M., Collie, A., Maruff, P., Darby, D. G., Bush, A., McCrory, P., & Bennell, K. (2001). Computerised cognitive assessment of concussed Australian Rules footballers. *British Journal of Sports Medicine, 35*(5), 354–360.
- Mallett, C. J., Kawabata, M., Newcombe, P., Otero-Forero, A., & Jackson, S. (2007). Sport Motivation Scale-6 (SMS-6): A revised six-factor sport motivation scale. *Psychology of Sport and Exercise, 8*, 600–614.

- Martens, R. (1975). *Social psychology and physical activity*. New York: Harper and Row.
- Martens, R., Landers, D. M., & Loy, J. W. (1972). *Sport cohesiveness questionnaire* (unpublished manuscript). Champaign, IL: University of Illinois.
- Maruff, P., Thomas, E., Cysique, L., Brew, B., Collie, A., Snyder, P., & Pietrzak, R. (2009). Validity of the CogState brief battery: Relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. *Archives of Clinical Neuropsychology*, 24(2), 165–178. Available from <https://doi.org/10.1093/arcln/acp010>.
- Matser, J. T., Kessels, A. G., Lezak, M. D., & Troost, J. (2001). A dose–response relation of headers and concussions with cognitive impairment in professional soccer players. *Journal of Clinical and Experimental Neuropsychology*, 23(6), 770–774.
- Mayers, L. B., & Redick, T. S. (2012). Clinical utility of ImPACT assessment for postconcussion return-to-play counseling: Psychometric issues. *Journal of Clinical and Experimental Neuropsychology*, 34(3), 235–242.
- 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(3), 261–270.
- McCrae, M. (2001a). Standardized mental status testing on the sideline after sport-related concussion. *Journal of Athletic Training (National Athletic Trainers' Association)*, 36 (3), 274–279.
- McCrae, M. (2001b). Standardized mental status testing on the sideline after sport-related concussion. *Journal of Athletic Training*, 36(3), 274–279.
- McCrae, M., Guskiewicz, K. M., Marshall, S. W., Barr, W., Randolph, C., Cantu, R. C., ... Kelly, J. P. (2003). Acute effects and recovery time following concussion in collegiate football players: The NCAA concussion study. *JAMA: The Journal of the American Medical Association*, 290(19), 2556–2563. Available from <https://doi.org/10.1001/jama.290.19.2556>.
- McCrae, M., Iverson, G. L., McAllister, T. W., Hammeke, T. A., Powell, M. R., Barr, W. B., & Kelly, J. P. (2009). An integrated review of recovery after mild traumatic brain injury (MTBI): implications for clinical management. *Clinical Neuropsychology*, 23, 1368–1390.
- McCrae, M., Kelly, J. P., Randolph, C., Cisler, R., & Berger, L. (2002). Immediate neurocognitive effects of concussion. *Neurosurgery*, 50(5), 1032–1042.
- McCrae, M. K., Kluge, J. P., Ackley, B., & Randolph, C. (1997). Standardized assessment of concussion in football players. *Neurology*, 48(3), 586–588.
- McCrae, M., Kelly, J. P., Randolph, C., Kluge, J., Bartolic, E., Finn, G., & Baxter, B. (1998). Standardized assessment of concussion (SAC): On-site mental status evaluation of the athlete. *The Journal of Head Trauma Rehabilitation*, 13(2), 27–35.
- McCrory, P. (2009). Sport concussion assessment tool 2. *Scandinavian Journal of Medicine & Science in Sports*, 19(3), 452.
- McCrory, P., Johnston, K., Meeuwisse, W., Aubry, M., Cantu, R., Dvorak, J., ... Schamasch, P. (2005). Summary and agreement statement of the second international conference on concussion in sport, Prague 2004. *Physician & Sportsmedicine*, 33(4), 29–44.
- McCrory, P., Meeuwisse, W., Aubry, M., Cantu, B., Dvorak, J., Echemendia, R., ... Turner, M. (2013). Consensus statement on concussion in sport—The 4th international conference on concussion in sport held in Zurich, November 2012. *Journal of Science & Medicine in Sport*, 16(3), 178–189. Available from <https://doi.org/10.1016/j.jsams.2013.02.009>.
- McCrory, P., Meeuwisse, W., Dvorak, J., Aubry, M., Bailes, J., Broglio, S. P., ... Voss, P. (2017). Consensus statement on concussion in sport—The 5th international conference on concussion in sport held in Berlin, October 2016. *British Journal of Sports Medicine*, 0, 1–10. Available from <https://doi.org/10.1136/bjsports-2017-097699>.

- McCrory, P., Meeuwisse, W., Johnston, K., Dvorak, J., Aubry, M., Molloy, M., & Cantu, R. (2009). Consensus statement on concussion in sport—The third international conference on concussion in sport held in Zurich, November 2008. *The Physician and Sportsmedicine*, 37(2), 141–159.
- McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Manual for the profile of mood states*. San Diego, CA: Educational and Industrial Testing Services.
- Moran, A. (2012). *Sport and exercise psychology: A critical introduction* (2nd ed.). New York: Routledge.
- Morgan, W. P., & Johnson, R. W. (1978). Personality characteristics of successful and unsuccessful oarsmen. *International Journal of Sports Psychology*, 9, 119–133.
- Morey, L. C. (2007). *Personality assessment inventory-adolescent (PAI-A)*. Lutz, FL: Psychological Assessment Resources.
- Moser, R. S., Iverson, G. L., Echemendia, R. J., Lovell, M. R., Schatz, P., Webbe, F. M., ... Silver, C. H. (2007). Neuropsychological evaluation in the diagnosis and management of sports-related concussion. *Archives of Clinical Neuropsychology*, 22(8), 909–916. Available from <https://doi.org/10.1016/j.acn.2007.09.004>.
- Moser, R. S., Schatz, P., Neidzwski, K., & Ott, S. D. (2011). Group versus individual administration affects baseline neurocognitive test performance. *American Journal of Sports Medicine*, 39(11), 2325–2330. Available from <https://doi.org/10.1177/0363546511417114>.
- Murphy, S. M. (1995). Transition in competitive sport: Maximizing individual potential. In S. M. Murphy (Ed.), *Sport psychology interventions* (pp. 341–346). Champaign, IL: Human Kinetics.
- Nia, M. E., & Besharat, M. A. (2010). Comparison of athletes' personality characteristics in individual and team sports. *Procedia-Social and Behavioral Sciences*, 5, 808–812.
- Omar-Fauzee, M. S., Daud, W. R. B., Abdullah, R., & Rashid, S. A. (2009). The effectiveness of imagery and coping strategies in sport performance. *European Journal of Social Sciences*, 9(1), 97–108.
- Pardini, J., & Collins, M. (2006). Creating a successful concussion management program at the high school level. In R. Echemendia (Ed.), *Sports neuropsychology: Assessment and management of traumatic brain injury* (pp. 142–159). New York: Guilford Press.
- Peterson, C. L., Ferrara, M. S., Mrazik, M., Piland, S., & Elliott, R. (2003). Evaluation of neuropsychological domain scores and postural stability following cerebral concussion in sports. *Clinical Journal of Sport Medicine : Official Journal of the Canadian Academy of Sport Medicine*, 13(4), 230–237.
- Philippe, R. A., Seiler, R., & Mengisen, W. (2004). Relationships of coping styles with type of sport. *Perceptual and Motor Skills*, 98(2), 479–486.
- Prapavessis, H. (2000). The POMS and sports performance: A review. *Journal of Applied Sport Psychology*, 12(1), 34–48.
- Public Broadcasting Service. (1997). *Alive on everest: RealAudio WebCast Transcript*. Retrieved from <http://www.pbs.org/wgbh/nova/everest/expeditions/97/summitaudiotrans.html>
- Raedke, T. D., & Smith, A. L. (2001). Development and preliminary validation of an athlete burnout measure. *Journal of Sport & Exercise Psychology*, 23, 281–306.
- Randolph, C., & Kirkwood, M. W. (2009). What are the real risks of sport-related concussion and are they modifiable? *Journal of the International Neuropsychological Society*, 15, 512–520.
- Randolph, C., McCea, M., & Barr, W. (2005). Is neuropsychological testing useful in the management of sport-related concussion? *Journal of Athletic Training*, 40(3), 139–152.

- Resch, J., Driscoll, A., McCaffrey, N., Brown, C., Ferrara, M. S., Macciocchi, S. N., ... Walpert, K. (2013). ImPACT test-retest reliability: Reliably unreliable? *Journal of Athletic Training*, 48(3), 506–511.
- Riemann, B. L., & Guskiewicz, K. M. (2000). Effects of mild head injury on postural stability as measured through clinical balance testing. *Journal of Athletic Training*, 35(1), 19–25.
- Robazza, C., Pellizzari, M., Bertollo, M., & Hanin, Y. L. (2008). Functional impact of emotions on athletic performance: Comparing the IZOF model and the directional perception approach. *Journal of Sports Sciences*, 26(10), 1033–1047.
- Rutter, M. (1985). Resilience in the face of adversity. Protective factors and resistance to psychiatric disorder. *The British Journal of Psychiatry*, 147(6), 598–611.
- Sarkar, M., & Fletcher, D. (2013). How should we measure psychological resilience in sport performers? *Measurement in Physical Education and Exercise Science*, 17(4), 264–280.
- Sarkar, M., & Fletcher, D. (2014). Psychological resilience in sport performers: A review of stressors and protective factors. *Journal of Sports Sciences*, 32, 1419–1434.
- Schatz, P. (2014). Computer instrumentation issues in sport-related concussion assessment. In R. Echemendia, & G. Iverson (Eds.), *The Oxford handbook of sports-related concussion*. Cambridge: Oxford University Press.
- Schatz, P., & Covassin, T. (2006). Neuropsychological testing programs for college athletes. In R. J. Echemendia (Ed.), *Sports neuropsychology: Assessment nad management of traumatic brain injury* (pp. 160–175). New York: Guilford Press.
- Schatz, P., Pardini, J., Lovell, M., Collins, M., & Podell, K. (2006). Sensitivity and specificity of the ImPACT test battery for concussion in athletes. *Archives of Clinical Neuropsychology*, 21(1), 91–99.
- Sheard, M., Golby, J., & van Wersch, A. (2009). Progress toward construct validation of the Sports Mental Toughness Questionnaire (SMTQ). *European Journal of Psychological Assessment*, 25(3), 186–193.
- Shehata, N., Wiley, J. P., Richea, S., Benson, B. W., Duits, L., & Meeuwisse, W. H. (2009). Sport concussion assessment tool: Baseline values for varsity collision sport athletes. *British Journal of Sports Medicine*, 43(10), 730–734.
- Silva, J. M., Shultz, B. B., Haslam, R. W., Martin, M. P., & Murray, D. F. (1985). Discriminating characteristics of contestants at the United States Olympic Wrestling Trials. *International Journal of Sport Psychology*, 16, 79–102.
- Smith, R. E., Schultz, R. W., Smoll, F. L., & Ptacek, J. T. (1995). Development and validation of a multidimensional measure of sport-specific psychological skills: The Athletic Coping Skills Inventory-28. *Journal of Sport & Exercise Psychology*, 17, 379–398.
- Snyder, A. R., & Bauer, R. M. (2014). A normative study of the sport concussion assessment tool (SCAT2) in children and adolescents. *The Clinical Neuropsychologist*, 28(7), 1091–1103.
- Spieler, M., Czech, D. R., Joyner, A. B., Munkasy, B., Gentner, N., & Long, J. (2007). Predicting athletic success: Factors contributing to the success of NCAA Division I AA collegiate football players. *Athletic Insight*, 9(2), 22–33.
- 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(9), 697–711.
- Thurman, D. J., Branche, C., & Snizek, J. (1998). The epidemiology of sports-related traumatic brain injuries in the United States: Recent developments. *The Journal of Head Trauma Rehabilitation*, 13, 1–8.
- Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *The American Journal of Psychology*, 9(4), 507–533.

- Valovich McLeod, T. C., Barr, W. B., McCrea, M., & Guskiewicz, K. M. (2006). Psychometric and measurement properties of concussion assessment tools in youth sports. *Journal of Athletic Training, 41*(4), 399–408.
- Valovich McLeod, T. C., Bay, R. C., Lam, K. C., & Chhabra, A. (2012). Representative baseline values on the Sport Concussion Assessment Tool 2 (SCAT2) in adolescent athletes vary by gender, grade, and concussion history. *American Journal of Sports Medicine, 40*(4), 927–933. Available from <https://doi.org/10.1177/0363546511431573>.
- Webbe, F. M., Salinas, C. M., Tiedemann, S. J., & Quackenbush, K. (2011). Personality: Contributions to performance, injury risk, and rehabilitation. In R. J. Echemendia, & C. T. Moorman (Eds.), *Praeger handbook of sports medicine and athlete health* (pp. 77–93). Santa Barbara, CA: Praeger.
- Weinberg, R. S., & Gould, D. (2014). *Foundations of sport and exercise psychology* (6th ed). Champaign, IL: Human Kinetics.
- Yukelson, D., Weinberg, R., & Jackson, A. (1984). A multidimensional group cohesion instrument for intercollegiate basketball teams. *Journal of Sport Psychology, 6*(1), 103–107.
- Zimmer, A., Marcinak, J., Hibyan, S., & Webbe, F. (2015). Normative values of major SCAT2 and SCAT3 components for a college athlete population. *Applied Neuropsychology. Adult, 22*(2), 132–140.

## Further reading

- Gaudreau, P., Blondin, J. P., & Lapierre, A. M. (2002). Athletes' coping during a competition: Relationship of coping strategies with positive affect, negative affect, and performance–goal discrepancy. *Psychology of Sport and Exercise, 3*(2), 125–150.
- Gibby, R. E., & Zickar, M. J. (2008). A history of the early days of personality testing in American industry: An obsession with adjustment. *History of Psychology, 11*(3), 164.
- Morgan, W. P. (1985). Selected psychological factors limiting performance: A mental health model. In D. H. Clarke, & H. M. Eckert (Eds.), *Limits of human performance* (pp. 70–80). Champaign, IL: Human Kinetics.
- National Collegiate Athletic Association (2014). Mental health best practices. [https://www.ncaa.org/sites/default/files/HS\\_Mental-Health-Best-Practices\\_20160317.pdf](https://www.ncaa.org/sites/default/files/HS_Mental-Health-Best-Practices_20160317.pdf).

## **Part VI**

# **Interviewing**

# Clinical interviewing

10

*Daniel N. Allen and Megan L. Becker*

Department of Psychology, University of Nevada, Las Vegas, NV, United States

## Introduction

Historically, the clinical interview has been a critical and often primary assessment component of mental health practice in psychology and psychiatry. As one of the most commonly used assessments, it allows for efficient and relevant information gathering for diagnostic and treatment considerations. Clinical interviews are also cost effective and efficient in comparison to other assessment methods that require specialized equipment (e.g., psychological and neuropsychological testing) and can be used across many settings and with a wide variety of patients with diverse presenting concerns. Interview style changed dramatically over the years but always involved a process whereby a clinician elicits a combination of social, medical, education, familial, psychological, developmental, and other information from the client and sometimes informants who know the client well (e.g., parent, spouse), either alone or within comprehensive psychological assessment. Information from other sources, such as medical and school records, may supplement information obtained from the client to establish diagnoses, judge the severity of symptoms, and prescribe treatment. The interview also provides an opportunity to build rapport with the client, which may be particularly important when the clinician intends to conduct additional assessment or treatment. The importance of the interview was evident in a survey of psychiatric practitioners that included 32 skills. 99.4% of respondents ranked the clinical interview as an important skill for psychiatrists and it was the highest ranked of all skills ([Langsley & Yager, 1988](#)).

Over the years, the clinical interview has become increasingly complex. This reflects changes in diagnostic systems, development of comprehensive assessment procedures to facilitate collection of more reliable and valid information, adaptation of interviews for specific settings and purposes, and changes to health care delivery systems. Aside from the variety of sources and information required by the clinical interview, the type of information gathered depends on setting, goals, and patient population. For example, different information is typically collected in outpatient mental health settings compared to inpatient medical settings, forensic interviews require more detailed and sometimes different information than what is gathered for a private practice diagnostic evaluation, and so on. Interpersonal factors influence the willingness of the client to provide accurate data and the clinician's ability to obtain that information. What is commonly referred to as "bedside manner" often

sets the tone for the interview and may affect the usefulness of the information obtained. Clinician biases, including those stemming from cultural factors that might impact the interview process, can introduce error into the interview process, thereby decreasing its validity. Client characteristics and motivation may also introduce error, such as in the case of a cognitively impaired individual who is unable to provide an accurate account of important historical details, or those involved in civil litigation where presence of a disorder would result in compensation. The current chapter is not intended as a comprehensive review of all aspects of clinical interviewing, as that is beyond the scope of this overview. However, this summary includes conceptual and historical background, description of some of the most common features and approaches to interviews, strengths and weaknesses of mainstream interview formats, and discussion of a number of critical topics (DSM-5, culture, technology) and their impact on the interview process.

## **Definition, history, and structure**

### ***Definition***

Broadly defined, [Wiens and Brazil \(2000\)](#) suggest that interviews are “face-to-face verbal exchanges in which one person, the interviewer, attempts to elicit information or expressions of opinion or belief from another person, the interviewee” (p. 389). Interviews are conducted for a variety of purposes. Some are conducted solely for informational purposes, such as commonly seen on television or radio talk shows where guests are interviewed by a host. The skilled host is capable of successfully eliciting information from guests with varied backgrounds and motivation/willingness to disclose information. Other interviews are conducted for employment purposes in which the goal is for the employer to determine whether a job applicant has the requisite experiences and qualifications for success in a job. This also offers the applicant an opportunity to acquire information considered in the decision to accept or reject a position if one were offered. Job interviews are often supplemented with other objective sources of information (psychological testing, work history, references, etc.) to improve successful hiring, as information obtained in the interview alone does not seem to provide an adequate basis for hiring decisions in most situations. Interviews may also be conducted for research purposes, where the goal is to gather consistent information across large groups of individuals as, for example, in political surveys. Research interviews are highly standardized with regard to interview content and structure, as well as sequencing of the interview questions. During the research interview, the interviewer typically reads a series of questions verbatim in a standard sequence in an identical manner to all interviewees, whose responses are recorded based on pre-coded choices.

In the context of this chapter we are concerned primarily with what is called the clinical interview. Many definitions of the clinical interview have been offered, to which we add the following one:

---

*The clinical interview is a distinct form of interviewing that involves a face-to-face verbal and nonverbal exchange between a clinician and client designed to gather data that is needed for diagnosis and treatment of the client.*

As suggested in this definition, the interview may be conducted for diagnostic or treatment purposes. However, the focus of the treatment interview is to help the client gain insight into their problems and develop strategies to address them, whereas the diagnostic interview involves gathering information that would allow for differential diagnosis and treatment prescription.

The definition also suggests that the clinical interview requires an exchange of information between two individuals and in this sense has a “conversational” aspect. However, simply having a conversation is not sufficient for the interview to be successful because an interview differs from a conversation in a number of important ways. First, the interview is a professional endeavor, often conducted by a mental or medical health professional who has been trained in the interview process and for a specific purpose related to the client’s psychological or physical well-being. To the extent that the interview requires complex or sensitive data, the training may be extensive and require refined clinical skill to complete well. Second, the type of information exchanged in the interview may differ substantially from the exchange that occurs in conversations. Both exchanges involve the verbal and nonverbal communication of objective and subjective information, and this information may vary widely from one interview to the next. However, unlike a conversation, the interview almost always involves topics that are considered sensitive or private in nature, such as trauma history, feelings of guilt, depression and suicidality, criminal behavior, sexual inadequacies and infidelities, medical information, etc. This type of information is not typically exchanged during a day-to-day conversation but is almost always the focus of the interview. Third, unlike conversations, information provided by the client in the interview is confidential in nature. There is a consent process that typically occurs prior to initiation of the interview where an “agreement” is formalized that outlines the limits of confidentiality and specific instances in which confidentiality may be broken by the clinician (e.g., danger to self or others). Potential legal and professional consequences exist for breaches in confidentiality that fall outside this agreement.

Fourth, the interview is always conducted to achieve a particular goal, which is not a requirement for a conversation. The interview goal varies, but generally involves considerations for diagnosis or treatment which are directly relevant to the client. Typically, the clinician and client mutually agree regarding the goal of the interview with the understanding that the content of the information discussed will be directly relevant to reaching that goal. Closely related to the interview goal, a fifth characteristic is that the interview is structured and planned to ensure that the goal is achieved. The extent of structuring varies and will be discussed in depth later in the chapter, but in contrast to conversations, interviews and questions are planned so as to elicit data directly related to this theme. Planning may include review of relevant psychological, medical, or academic records prior to the interview, with specific questions identified to fill in gaps that are apparent after record

review. Interview length, meeting place and time, and other similar features are also planned. In this regard, interviews may lack the spontaneity of conversations.

A sixth distinction involves the roles of the clinician and client during the interview. Throughout the interview, the clinician is typically responsible for directing the course of the conversation to meet the interview goal while the client follows the lead of the clinician, providing information that is directly relevant to the questions posed by the clinician. As such, the client typically provides most of the information exchanged in the interview, talking approximately 80% of the interview time. While both maintain some degree of control during the interview, the roles are nonreciprocal in the sense that the goal of the interview is to provide benefit to the client, not the clinician.

There are other important distinctions between a conversation and an interview that could be discussed, but those mentioned above provide ideas about some of the major differences between a conversation and interview. The structure of the clinical interview selected may influence the extent to which each of these aspects impacts the interview. Later sections of this chapter provide detailed discussion of considerations relating to interview structure, but before that we provide a brief review of the historical development of the clinical interview.

## ***History of the clinical interview***

The history of clinical interviewing coursed through the mid-20th century and was built upon foundational theories driven by physicians and philosophers in the early 1900s. In the previous edition of this Handbook, [Shea \(2000\)](#) provided a detailed review of this history, much of which is summarized in [Table 10.1](#). In this chapter, we highlight some selected important events and provide an update of developments since then. At the turn of the 20th century, Kraepelin classified mental illness into manic depression and dementia praecox ([Kaplan, Freedman, & Sadock, 1980](#)). This important and enduring classification set the stage for development of more detailed criterion sets that proved important in efforts to structure the clinical interview. Decades later, Heinz Hartman and Anna Freud contemplated the role of defense mechanisms in the clinical interview which aided in forming the patient's clinical picture ([Freud, 1936; Hartmann, 1939](#)). During this time, clinical interviews spanned across multiple days as clinicians carefully constructed theories about the patient's functioning, as no formal diagnostic classification system existed to inform treatment other than the rudimentary distinctions drawn by those like Kreapelin. Adolf Meyer developed the first semistructured psychiatric interview in 1951 and coined the term "mental status" as he articulated the importance of obtaining a patient history to contextualize the presenting concern. Following this in 1954, Harry Stack Sullivan highlighted how the interviewer's behavior could influence that of the patient and thereby distort symptomology ([Sullivan, 1970](#)). Carl Rogers introduced his concepts "client-centered approach" and "unconditional positive regard" and integrated these into clinical interviews stating that interviewers must use a tactful and socially naturalistic interviewing style to foster human connection ([Rogers, 1951, 1959](#)). Concurrently and into the 1960s, Karl Jaspers and Medard Boss described a way of

**Table 10.1** Important historical events in the evolution of the clinical interview

Year	Historical contribution	Reference
Early 1900s	<ul style="list-style-type: none"> <li>Kraepelin—classified mental illnesses</li> <li>Differentiated manic depression from dementia praecox</li> </ul>	Kaplan et al. (1980)
1936, 1939	<ul style="list-style-type: none"> <li>Heinz Hartman and Anna Freud—ego psychology and defense mechanisms</li> <li>Patient's defenses manifested in the interview</li> </ul>	Freud (1936) and Hartmann (1939)
1944	<ul style="list-style-type: none"> <li>J.C. Whitehorn—eliciting patient opinion</li> <li>Exploration of opinions about interpersonal relations and reasons for caring for others</li> </ul>	Whitehorn (1944)
1951	<ul style="list-style-type: none"> <li>Adolf Meyer—development of semistructured psychiatric interview</li> <li>Patient biography and influences on the patient's current behavior</li> <li>Defined term “mental status”</li> </ul>	Donnelly, Rosenberg, and Fleeson (1970), Meyer (1951)
1954	<ul style="list-style-type: none"> <li>Harry Stack Sullivan—interviewer's behavior impacts the patient's, with potential for distorting symptomology</li> <li>Flexibly structured style of interviewing</li> </ul>	Sullivan (1970)
1951, 1959	<ul style="list-style-type: none"> <li>Carl Rogers—“client-centered approach”</li> <li>“Unconditional positive regard”</li> <li>Interviewers use a naturalistic sense their social skills and personality</li> </ul>	Rogers (1951, 1959)
1952	<ul style="list-style-type: none"> <li>Theodore Reik—interviewing about free-floating attention, conscious and unconscious observation, and the therapist–patient alliance</li> </ul>	Reik (1952)
1954	<ul style="list-style-type: none"> <li>Newman and Redlich—patient/physician dialogue</li> </ul>	Gill, Newman, and Redlich (1954)
1955	<ul style="list-style-type: none"> <li>Felix Deutsch and William Murphy—associative anamnesis which emphasizes free association and gentle probing</li> </ul>	Deutsch and Murphy (1955a, 1955b)

(Continued)

**Table 10.1** (Continued)

<b>Year</b>	<b>Historical contribution</b>	<b>Reference</b>
1950s and 1960s	<ul style="list-style-type: none"> <li>Karl Jaspers and Medard Boss—understanding patient's experience “being in the world”</li> <li>Delicately probing the patient for symptoms, feelings, perceptions, and opinions</li> </ul>	Hall and Lindzey (1978)
1965	<ul style="list-style-type: none"> <li>Richardson, Dohrenwend, and Klein—characteristics of the interview process such as style of questioning (open- or closed-ended)</li> </ul>	Richardson et al. (1965)
1969	<ul style="list-style-type: none"> <li>Alfred Benjamin—genuineness and common sense in the therapeutic relationship</li> <li>Trusting relationship with the patient, clinician does not hide behind rules, position, or authority</li> </ul>	Benjamin (1969)
1971	<ul style="list-style-type: none"> <li>MacKinnon and Michels—analytic and dynamic principles in the initial interview and therapy</li> <li>Exploration of patient defense mechanisms and clinician style</li> </ul>	MacKinnon and Michels (1971)
1972	<ul style="list-style-type: none"> <li>Feighner criteria—15 diagnostic categories by using both exclusion and inclusion criteria</li> </ul>	Feighner et al. (1972)
1975	<ul style="list-style-type: none"> <li>Egan—describing information in a concrete language to convey ideas in an educational sense</li> </ul>	Egan (1975)
1975	<ul style="list-style-type: none"> <li>Grinder and Bandler—phrasing questions so that the patient's hidden thoughts are gradually pulled to the surface</li> </ul>	Grinder and Bandler (1975)
1978	<ul style="list-style-type: none"> <li>Spitzer, Endicott, and Robins—Research Diagnostic Criteria (RDC) which added 23 disorders to psychopathological range</li> </ul>	Spitzer et al. (1978)
1978	<ul style="list-style-type: none"> <li>Endicott and Spitzer—Schedule for Affective Disorders and Schizophrenia (SAD)</li> </ul>	Endicott and Spitzer (1978)
1978	<ul style="list-style-type: none"> <li>Alfred Margulies and Leston Havens—importance of a phenomenological approach</li> <li>Counterprojective technique</li> </ul>	Havens (1978, 1979), Margulies (1984, 1981)

(Continued)

**Table 10.1** (Continued)

Year	Historical contribution	Reference
1981	<ul style="list-style-type: none"> <li>• Robins—Diagnostic Interview Schedule (DIS) used by lay interviewers and highly scheduled to ensure interrater reliability</li> </ul>	Robins et al. (1981)
1983	<ul style="list-style-type: none"> <li>• Spitzer and Williams—Structured Clinical Interview for the DSM-III (the SCID-III, SCID-III-R, and the SCID-IV)</li> </ul>	Spitzer and Williams (1983)
1983	<ul style="list-style-type: none"> <li>• Gerald Pascal—behavioral incident</li> <li>• Questions can range on a continuum from a request for patient opinion to historical or behavioral description</li> </ul>	Pascal (1983)
1983	<ul style="list-style-type: none"> <li>• William Miller outlines motivational interviewing</li> </ul>	Rollnick and Miller (1995)
1985	<ul style="list-style-type: none"> <li>• Hersen and Turner—interviewers should be familiar with specific techniques for exploring diagnostic criteria from the various diagnoses in the DSM-III</li> </ul>	Hersen and Turner (1985)
1988	<ul style="list-style-type: none"> <li>• Shea, 1988—streamlined interviewing knowledge</li> <li>• synthesized wide range of important information into an interview in a natural and dynamic way</li> </ul>	Shea (1988a, 1988b)
1989	<ul style="list-style-type: none"> <li>• Othmer and Othmer—practical tips and model questions for interviewing with DSM-III-R interview situations involving complicated clinical interactions</li> </ul>	Othmer and Othmer (1989, 1994a, 1994b)
1998–1999	<ul style="list-style-type: none"> <li>• Shea—facilitic principles applied to interviewing about suicidal ideation</li> </ul>	Shea (1988a, 1988b, 1999)
1999	<ul style="list-style-type: none"> <li>• Lecrubier—the Mini-International Neuropsychiatric Interview (MINI)</li> <li>• Pioneered in the United States by Sheehan</li> </ul>	Lecrubier et al. (1998)
2012	<ul style="list-style-type: none"> <li>• Movement toward evidence-based clinical interviewing</li> </ul>	Henderson et al. (2012), Hersen and Sturmey (2012), Saywitz and Camparo (2013)

probing for the patient's symptoms, feelings, perceptions, and opinions with questions which led to further discussion about question style such as open- or closed-ended (Hall & Lindzey, 1978; Richardson, Dohrenwend, & Klein, 1965). Later, the Feighner criteria introduced 15 diagnoses that delineated diagnostic inclusion and exclusion criteria in 1972 (Feighner et al., 1972). Six years later, 23 disorders were added from the Research Diagnostic Criteria (RDC) (Spitzer, Endicott, & Robins, 1978). The Diagnostic Interview Schedule (DIS) appeared in 1981 just prior to the Structured Clinical Interview for the DSM-III (Robins, Helzer, Croughan, & Ratcliff, 1981; Spitzer & Williams, 1983) which provided a standardized, yet flexible way to assess for symptoms that met diagnostic criteria. With an abundance of available information, Shawn Shae synthesized diagnostic information into interviews in a natural and dynamic way allowing for improved conversational flow (Shae, 1988). Subsequent editions of these semistructured interviews evolved over the 1990s with the advent of the DSM-IV in 1994 and the DSM-5 in 2013 (American Psychiatric Association, 1994, 2013).

In recent years, clinicians placed a greater emphasis on evidence-based clinical interviews. Diagnostic classifications between versions of the DSM were formulated based on research. However, increased attention was granted toward evidence-based interviewing strategies and approaches to differential diagnosis. This focus resulted from attempts to improve diagnostic accuracy by understanding the complexities of clinical judgement and differences in decision making. For example, a novice clinician may gravitate toward ruling out a greater number of possible diagnoses, but experienced clinicians may be more prone to rely on their experience; each of which result in separate consequences (Henderson, Tierney, & Smetana, 2012). Although evidenced-based diagnostic classification and clinical decision making improve reliability and validity of diagnosis overall, they remain subject to methodological limitations and there is difficulty in applying nomothetic conclusions to ideographic situations in clinical practice.

Motivational interviewing is a type of evidence-based clinical interview outlined by William Miller in 1983 which is aimed at gauging the patient's motivation to change a behavior. Its use has grown substantially over the past 20 years. Its brief implementation lent it favorability in integrated healthcare settings for patients with various psychological and/or medical conditions such as substance use, poor diet, and lack of exercise, among others. Motivational interviewing emphasizes the collaborative process of the patient and interviewer in understanding the patient's readiness to change a behavior and how important the patient perceives it is to change. The clinician and patient then reflect on the benefits and problems with continuing the current behavior juxtaposed against those of the alternative (healthier) behavior and depending on the patient's goal, problem solve the steps necessary to promote change. In summary, motivational interviewing presents an interesting direction for clinical interviewing as mental health integrated into primary care and other medical settings. The historical progression of clinical interviewing changed from lengthy multi-day interviews to cumbersome multi-hour interviews which are continuously refined to meet the practical demands of mental health care as it integrates into contemporary holistic models of care.

## **Structure of the clinical interview**

The clinical interview is directed toward accomplishing a specific goal, although clinicians may employ substantial variability in their approaches. Much of this variability can be understood by examining elements that establish structure of the interview. In this regard, Richardson et al. (1965) draw a distinction between *standardization* and *scheduling* aspects of the clinical interview. *Standardization* addresses the informational areas covered in the interview process, while *scheduling* addresses the degree to which the wording and sequence of an interview is specified beforehand. These key concepts provide a framework for the four main types of modern interview procedures: free format or open, flexibly structured, semistructured, and fully structured interviews. Additional consideration of the advantages and disadvantages of each of these will be discussed later in this chapter. Regarding applying the standardization and scheduling concepts to these four types of interviews, formal standardization and scheduling are absent from the free-format interview. The clinician is free to cover any informational area deemed relevant to the presenting concern and scheduling may be unique for each client. In sharp contrast, the fully structured interview relies on a highly standardized and schedule format. Areas of information covered by the interview are specified, order of information coverage is fixed, and content of interview questions is stipulated. The flexibly structured and semistructured interviews fall between free format and fully structured interviews with regard to standardization and scheduling. Both are standardized in that the information areas to be covered are specified, although there is flexibility in scheduling aspects. The flexibly structured interview might be considered closer to the free format interview as the clinician obtains the standardized information using whatever order and question content seems best based on the client's unique characteristics and presenting concerns. In fact, the flexibly structured interview often begins in a free format style, where topics pressing to the client are initially explored, and the standardized information is obtained afterward. The semi-structured interview imposes more structure on scheduling with order of information often specified and some question content also stipulated. However, depending on the interview, there is greater flexibility for the clinician to move within these scheduling guidelines when eliciting information from the client.

Shea (2000) identifies other important structural factors that determine interview style, including: "(1) specific content areas required to make a clinical decision or to satisfy a research data base, (2) quantity of data required, (3) importance placed on acquiring valid historical and symptomatic data as opposed to patient opinion and psychodynamic understanding, and (4) time constraints placed upon the interviewer" (p. 340). As previously discussed, the specific content areas (1) and quantity of data required (2) are addressed based on standardization and scheduling aspects (Richardson et al., 1965) which together describe interview types that range from completely free-form and spontaneous to highly structured that constrain the clinician, organization, and content.

To the extent that the historical data (3) is a vital component in support of the goal of the interview, greater attention is devoted to ensuring the accuracy of the

patient report. This would be particularly important when, for example, the goal of the interview is to establish a diagnosis in which symptoms required for differential diagnosis may (1) not be present at the time of the interview itself or (2) are required to occur over a long period of time. Differential diagnosis between schizophrenia, schizoaffective disorder, and a mood disorder with psychotic features is an example of the former. It poses a particular diagnostic challenge in many cases, because the presence of past psychotic symptoms must be ruled out in clients exhibiting depressive symptoms at the time of the interview before a definitive mood diagnosis is assigned. A similar challenge exists for clients presenting with psychotic symptoms. History and duration of manic or depressive episodes must be identified to establish the differential diagnosis of schizophrenia versus schizoaffective disorders. Persistent depressive disorder (*dysthymia*) also can be diagnostically challenging because it requires documenting the presence of depressive symptoms that are present more days than not for a period of two years. Diagnosis is further complicated by the presence of symptoms that are consistent with major depressive disorder. In these cases, establishing memorable historical events (e.g., winter break of freshman year of college) or focusing on time periods when the symptoms are most likely to meet diagnostic criteria (e.g., “When were the symptoms the worst? Let’s talk about that time.”) may be helpful. In other instances, informants may be interviewed. Informant reports of historical information are useful for children or adults who are unable to provide accurate data due to their age, presence of cognitive disorder, etc., when there is a concern regarding the truthfulness of the client, or when required by the goal of the assessment itself. However, the clinician should keep in mind that informants may have their own motivation or biases when reporting information that can affect veracity of the reports.

Finally, time constraints (4) have become an increasing concern in the era of managed care, where most intake interviews are limited to one hour or less. Needless to say, this move has resulted in much more focused and specialized interviews, where the goal of “getting to know the client” is subordinated to other superordinate goals. While this is common practice currently, much longer periods of time were traditionally devoted to the initial interview. In fact, as previously mentioned, in the first quarter of the 20th century when interviewing methods were first developing in psychiatry, clinicians could devote many hours over many days to complete the initial interview (Shea, 2000). Interviews were largely unstructured since there were not clear criteria for mental disorders in the absence of a well-developed diagnostic system, and there was a corresponding lack of interventions that were targeted for specific mental disorders. With the development of the DSM system and corresponding treatments designed for the specific DSM disorders, the need for increased standardization and efficiency have grown substantially so that a diagnosis can be made quickly followed by prescription for the appropriate treatment. The next section briefly discusses reliability and validity issues relevant to clinical interviews and provides information regarding the advantages and disadvantages of current mainstream interview formats (free format or open, flexibly structured, semistructured, and fully structured interviews) to provide the reader with information useful in selecting an interview format.

## Selecting an interview format

### ***Reliability and validity***

Currently, debate regarding the most appropriate clinical interview format mainly revolves around psychometric issues including reliability and validity of diagnosis. Historically, diagnostic reliability for the clinical interview has proven challenging. Indeed, studies have reported widespread diagnostic disagreement between clinicians and biases in diagnoses seemingly related to individual characteristics, such as race of the client. Main sources of error that contribute to unreliable diagnoses result from variability in: (1) *criterion* used for diagnosis, (2) *information* used to make a diagnosis, and (3) *client characteristics* (Spitzer et al., 1975; Ward, Beck, Mendelson, Mock, & Erbaugh, 1962).

*Criterion* variance arises when clinicians use different diagnostic criteria and thresholds to make diagnoses. Certainly, the use of different diagnostic systems (RDC vs DSM) will lead to differences in diagnosis but with the widespread adoption of the DSM system, this source of variance is almost completely eliminated. However, clinicians may vary in how they implement the DSM system by using different standards to determine the presence and severity of symptoms or using different standards to determine whether the symptoms are clinically significant (the degree to which the symptoms impair functioning of or are distressing to the client). For example, determining if the psychological distress associated with anxiety symptoms is severe enough to warrant a diagnosis of post-traumatic stress disorder or functional consequences of substance use are severe enough to warrant a substance use disorder diagnosis, is a subjective process that may be influenced by a number of factors and will certainly influence diagnosis.

*Information* variance results when clinicians consider different sources of information for the client. Self-report versus informant report is an obvious example of two different sources of information that may provide different information about the client, leading to different diagnoses. However, within the context of self-report (or informant report), clinicians may use different questions to elicit information, observe and emphasize different behaviors as clinically meaningful, and organize information differently, all which may affect reliability of diagnosis.

*Client* variance occurs when differences within the same client cause differences in clinical presentation. Clients may exhibit different diagnosis at different time points, referred to as *subject* variance (Spitzer et al., 1975), such as a client who initially presents with a depressive disorder and then later develops an alcohol use disorder. Temporal symptom fluctuations in the same patient who is observed at different time-points (e.g., bipolar depressed vs manic) may also lead to significant discrepancies in clinical presentation and subsequent diagnosis, referred to as *occasion* variance.

Of these sources of variance that contribute to unreliable diagnoses, criterion variance has historically accounted for most diagnostic variability (Spitzer et al., 1975). In fact, Ward et al. (1962) suggested that 62.5% of diagnostic unreliability was accounted for by criterion variance, 32.5% by information variance, and only 5.0% from patient variance. Some types of interviews are more susceptible to

influence by these sources of variance. [Blashfield \(1992\)](#) indicates that use of unstructured interviews results in failure to systematically apply diagnostic criteria (criterion variance) resulting in misdiagnoses of approximately 60% of the patients. With the broad acceptance and continued development of DSM criteria, including Spitzer's work with the DSM-III, it is expected that criterion variance has decreased substantially since [Ward et al. \(1962\)](#) and Spitzer et al. (1975) published their statements. However, common sources of information variance continue to contribute to unreliability in diagnosis. [Rogers \(2001\)](#) suggests that common sources of needless variability include idiosyncratic questions, idiosyncratic content coverage, idiosyncratic sequencing of diagnostic questions, idiosyncratic recording of information, and absence of formal ratings to establish severity of symptoms. Many of these concerns may be addressed through standardization and scheduling. Defining the areas to be assessed, specifying the order and wording of questions, and systematizing how client responses are rated addresses many of the concerns regarding criterion and information variance inherent within the clinical interview. Selection of an interview format that is capable of addressing the goals of the clinical interview and also minimizes criterion and information variance is optimal, and the following sections address the benefits and shortcomings of the current four main interview styles.

### ***Free-format or open interviews***

Open or free-format clinical interviews allow clinicians to conduct the interview however they see fit ([Akin & Turner, 2006](#)) and are widely used in clinical settings. The interview often takes a conversational tone between the clinician and client. Standardization and scheduling aspects of the interview vary greatly from one clinician to the next. The clinician determines the content areas covered, level of detail in which each area is covered, order in which the information is collected, questions asked to elicit information, length of the interview, manner in which information is rated, recorded and interpreted, and any other aspects deemed relevant.

Some advantages of this method are that the interview can be tailored to the specific concerns of the client, it allows flexibility in how much depth a presenting problem or symptom will be explored, the conversational nature can help build rapport, it may be less time consuming than other interview types, there are no specialized forms for recording information, and it can be administered anywhere. The clinician can exercise discretion regarding the appropriateness of inquiring about specific information at a specific point in time based on the client's reactions because the clinician is allowed to determine the order in which information is obtained, reword questions as needed, and pose new questions when indicated. Thus, the obtained information may be more descriptive of the client's day to day experiences, reactions, and feelings. Clinicians may also value this type of interview for its capability of uncovering client feelings, defenses, and thought processes which provides insight for psychotherapy.

Criticisms of the free-format interview have typically focused on threats to reliability and validity because there is wide variability in the type of information

gathered from one interview to the next based on the clinician's theoretical orientation and the presenting complaints of the client. Interview content is determined by the clinician, so the questions asked may vary according to the clinician's bias or theoretical perspective. Consequently, the interview may focus on information that is trivial in nature at the expense of omitting relevant information important for accurate diagnosis. Additionally, other uncontrolled sources of variation introduce substantial variance into data collection during the interview, which in turn negatively impacts the reliability and validity of diagnoses. [Edelbrock and Bohner \(2000\)](#) describe a simple experiment to illustrate the potential pitfalls of the free format-interview:

*Suppose there was a pool of subjects who were absolutely identical in every way. . . . Different interviewers would ask different questions in different ways, and. . . . interviews conducted by the same interviewer might yield quite different information due to variations in interviewing style and content. . . [In] this hypothetical example. . . . subject variance would be eliminated, since all subjects are identical. The criterion variance. . . could also be eliminated if one diagnostic system were used. But information variance would remain as a major threat to reliability (pp. 369–70).*

From a measurement perspective, the free-format interview's lack of formal structure makes it subject to error. Due to its unstructured nature, the free-format interview also requires an extensive conceptual understanding of the diagnostic system, and more experience, skill, and training is required for this type of interview compared to more standardized interview formats.

### **Flexibly structured interviews**

Flexibly structured interviews are the most popular type of clinical interview used by experienced clinicians. They are flexible because standardization and scheduling aspects of the interview are ultimately determined by the clinician and may vary from one clinician to the next. However, structure is also imposed by the clinician on the data obtained in the interview and to a lesser degree, the order in which the information is gathered. The flexibly structured interview typically begins with a free-format style where the information discussed is primarily determined by those issues that are most relevant or serious for the client (often times the presenting concern) which helps to increase engagement and establish rapport. Once rapport is established, the clinician begins to acquire information during the interview.

From a standardization perspective, the information collected during the interview may be significantly influenced by the goal of the interview and the training and theoretical orientation of the clinician. The interview may appear spontaneous to the casual observer with little consistency from one clinician to the next, but careful evaluation reveals topics that are consistently covered and similar questions are used to elicit information across interviewers and clients. Typically, flexibly structured interviews begin with questions about clients presenting concern or chief

complaint and determine basic demographic and identifying information (e.g., age, years of education, marital status, race/ethnicity, occupation, residence, referral information). Questions that address the history of the presenting concern include duration and severity of current symptoms, psychosocial and other stressors that precipitated current problems, and ways that the problem has impacted functioning at home, work, and socially. Importantly for diagnostic purposes, questions are included regarding the presence and severity of current psychiatric symptoms. Developmental, social, family, and medical history are used to supplement diagnostic information and present a depiction of the client independently of the disorder. Developmental information may include educational history and any learning difficulties, trauma, family constellation, divorce, etc. Historical information is obtained for prior medical and mental disorders and for treatment, and so is information about current medications, alcohol and drug use, and disabilities. Family history of mental and medical disorders, occupational history, legal involvement, military service, and other information may also be collected.

The flexibly structured interview also typically includes an evaluation of mental status, which is an objective description of the client's appearance, behavior, speech, mood and affect, thought content and process, insight, and cognition at the time of the evaluation. There is some variability among clinicians in the procedures used to assess these areas, but informal observation during the interview and the careful assessment of thought processes based on client speech are essential. Informal observations of behavior and speech provide information about client appearance, grooming, affect, motor activity, sequencing, of thoughts and distractibility. Specific questions are also included that focus on various psychiatric symptoms such as delusions and hallucinations. Brief formalized procedures to determine orientation (e.g., day, date, month, year, place) and cognitive abilities such as attention, memory, and abstraction are typically included. The mental status examination may be lengthy or brief depending on the purposes of the interview. Regardless, its purpose is to provide an accurate description of the client's mental state at the time of the interview that will be useful for diagnostic and treatment purposes, including determinations about insight into present illness and judgment.

Finally, the clinician writes a formal report that contains a synthesis of supplemental background and historical data and a summary of the results of the clinical interview. This often includes a recommendation regarding symptoms that should be treated and the type of treatment that is most suitable for the client.

Another notable strength of the flexibly structured interview is that it is easily modified by clinicians to address numerous diagnostic and treatment goals within the context of external constraints placed on the interview, such as the amount of time allocated for the interview. This allows for clinicians to evaluate either a narrow or wide range of symptoms and behaviors depending on the goal of the interview. Clinicians may be less likely to miss important diagnoses or treatment considerations in comparison the highly structured and standardized interview approaches with strict standardization of the information coverage. However, like free-format interviews, flexibly structured interviews have been criticized due to a lack of formal structure which increases potential influence of criterion,

information, and subject variance on final diagnosis. Also, the flexibly structured interview requires an extensive conceptual understanding of the diagnostic system. Thus, more experience, skill, and training is required for this type of interview compared to some of the other more standardized interview formats.

### ***Structured and semistructured interviews***

As noted in the previous sections, the clinical interview is widely used because it is flexible and adaptable, typically requires no specialized equipment, allows for gathering information on a broad range of phenomenon, is cost effective, and is generally well received by both clinicians and clients. However, as also noted, it is prone error from criterion, information, and subject variance, which if not controlled results in unreliable observations and diagnoses that are unlikely to be useful in prevention efforts, prescribing treatment, or stimulating research. One proposed solution to this problem is standardizing and scheduling the interview, as this involves clearly defining the phenomena assessed, prescribing the questions and their order, and providing the same methods to rate, record, and interpret client responses. These procedures reduce error from criterion and information variance, which has contributed to the growing popularity and development of structured and semistructured interviews.

The movement toward more structured approaches spurred forward with the publication of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III) ([APA, 1980](#)) which was accepted by many clinicians and researchers at the time of its publication. Although not without limitations, it provided the first unified and detailed criteria sets for diagnosis. The need for methods to gather uniform and detailed data to support differential diagnoses also grew, because diagnostic criteria became more well established with the publication of the DSM-III. In response to this need, [Spitzer and colleagues \(1978\)](#) developed the Research Diagnostic Criteria (RDC) which provided specific criteria for many functional disorders, followed by the publication of other structured and semistructured interviews such as the Schedule for Affective Disorders ([Endicott & Spitzer, 1978](#)) and the Diagnostic Interview Schedule (DIS; [Robins, Cottler, Bucholz, & Compton, 1995](#)) (the DIS is reviewed in the Diagnostic and Symptom Interviews for Adults chapter in this handbook). Similar interviews in addition to these became gold standards for diagnosis in research settings, although clinicians were less enthusiastic about using them in clinical practice. There is an ongoing need to develop structured and semistructured interviews that are simpler and easier to navigate so the advantages afforded by these interviews can be realized in clinical practice.

***Fully structured interview.*** Fully structured interviews provide detailed specification about the type of information included in the interview and how it is collected. In this sense, the fully structured interview is fundamentally a list of target symptoms and behaviors, with clear detailed guidelines that govern how the interview is performed and how the data are recorded ([Edelbrock & Costello, 1984](#)). Edelbrock and Contello concluded that interviewers may be viewed as interchangeable pieces of the assessment machinery, because the wording of questions, the sequence of

questions, and recording and rating responses are all specified. In these interviews, little clinical inference is required while collecting the data and there is no need for clinical interpretation in the diagnostic process, which ensures that different interviewers will obtain the same data on the same patient. Minimizing the need for clinical judgment is especially important for interviews like the DIS that are designed for lay interviewers, but it also helps mitigate influence of diagnostic biases clinicians may have.

Some advantages of structured interviews include that: (1) they reduce unnecessary error resulting from criterion and information variance; (2) the systematized evaluation of symptoms and behavior is likely to reduce misdiagnosis; (3) when they are comprehensive there is less likelihood of missing a diagnosis since diagnostic categories are not arbitrarily excluded based on the clinician's assumptions or working hypotheses; (4) they ensure a high level of standardization; and (5) they increase confidence for systematic comparisons made across settings, time, and diagnosis within the same client (as often occurs in clinical settings) and between different clients (as often occurs in research settings) (Rogers, 2001). These qualities are responsible for making structured interviews the instrument of choice in research settings.

However, disadvantages include: (1) advanced specialized training is usually required to ensure reliable administration and scoring; (2) many tend to be quite extensive and time consuming to complete, rendering them impractical in some clinical settings; (3) they allow limited flexibility in gathering information based on the unique characteristics and concerns of the clients; (4) they are not well equipped to address threats to validity of data due to resistance, dissimulation, etc.; and (5) they cannot account for all possible diagnosis. With regard to this latter point, some suggest that in their attempt to minimize information variance, structured interviews may increase criterion variance (Rogers, 2001). Despite evidence supporting the reliability of structured interviews, it is also important to recognize that their reliability should not be assumed. Monitoring reliability of clinicians who complete these interviews helps guard against rater drift and other factors that negatively impact diagnostic reliability.

*Semistructured interviews.* Semistructured interviews differ from fully structured interviews in that they allow clinicians more flexibility during the interview and ensure that a certain set of questions are covered in the interview (Akin & Turner, 2006). Although the coverage of certain areas are specified, the clinician has significant latitude in the sequencing of the information. Questions are provided to elicit data, but the clinician can also determine their wording. For example, the interview may start by addressing the client's presenting problem followed by relevant historical information, illness episodes, etc., but the clinician may choose to deviate from this suggested sequence if clinically warranted. In some interviews, standardized questions are required, but the clinician determines whether the information in the client's response is sufficient to meet diagnostic criteria. If it remains unclear whether the response meets diagnostic criteria, the clinician follows up with unscripted questions to establish the presence or absence of the criteria. Semistructured interviews are perceived as more natural than structured interviews because they can be

personalized to the client and allow for more spontaneity, although greater clinical judgment is required compared to structured interviews (Edelbrock & Costello, 1984). The SCID is an excellent example of how clinical judgment is required for semistructured interviews. In the training materials, the SCID developers emphasize the “C” or clinical judgment aspect of the SCID, indicating that if an interviewer is not able to conduct a competent clinical interview without the SCID, they can’t conduct one with the SCID.

Semistructured interviews share some of the same advantages of structured interviews regarding standardization and reducing diagnostic error associated with criterion and information variance. As indicated, they also permit examiners to ask questions in a manner that is consistent with their clinical style and based on the unique characteristics and concerns of the client. This includes formulating their own questions to clarify presence or absence of diagnostic criteria and rephrasing questions when it appears that the client does not understand. The interview is more natural and spontaneous than a fully structured interview (Akin & Turner, 2006; Rogers, 2001). Comprehensive interviews like the SCID ensure systematic and thorough coverage of symptoms and diagnoses, to assure meeting of specific criteria and consideration of all diagnoses. Some interviews also allow clinicians the liberty to explore diagnoses that are not included in the interview itself, when clinically indicated.

Like structured interviews, disadvantages include that semistructured interviews are time consuming, require extensive training, and manuals, workbooks, and/or scoring sheets are required for proper administration. Although expected to decrease criterion variance, semistructured interviews may potentially increase information variance.

## Critical topics

### ***Culture and diversity***

Psychiatric symptoms may not manifest in an identical manner across various cultures or may have different cultural meanings (Lewis-Fernández et al., 2010). This observation led to increased consideration of race and ethnicity as an important component of the clinical interview to minimize error attributable to examiner bias and patient variance. The Diagnostic and Statistical Manual of Mental Disorders acknowledged the importance of cultural differences in diagnosis and included resources in its most recent version (DSM-5) that are designed to help clinicians increase consideration of cultural factors. The DSM-5 includes available background information for each diagnosis that discusses race, gender, and other factors relevant to diagnosis. It also contains a glossary of cultural concepts of distress, an updated cultural formulation outline, a Cultural Formulation Interview, and some cultural based disorders are included as V codes (APA, 2013). Research reporting importance and influence of cultural and ethnic factors in psychological assessment continues to grow and the importance of cultural considerations in the clinical

interview is discussed (Adebimpe, 1994; Dana, 2008; Al-issa, 1995; Neighbors et al., 1989; Paniagua, 2001; Trierweiler, Neighbors, Munday, Thompson, & Binion, 2000). Clinicians should consider the impact that culture may have on the meaning and significance of the information, because the clinical interview is an interpersonal interaction where information is gathered. Clinicians who have increased their cultural sensitivity through formal training, self-education, consultations, cultural humility, and other methods, are most capable of conducting effective clinical interviews with those from other cultures and are viewed as more competent than clinicians who are blind to aspects of the client's culture (Pomales et al., 1986; Want et al., 2004). Thus, considering how the client's clinical picture is influenced by culture both independently and through intersectionality can affect the client's experience with mental health care, appropriately adjust the information discussed in the interview, and improve diagnostic accuracy.

Zink, Lee, and Allen (2015) discussed numerous cultural factors to consider in the clinical interview including socio-economic status and population characteristics; racial identity; racism, discrimination, and stereotypes; language; cultural mistrust; and diagnostic bias. For more detail on these issues, the reader is encouraged to review Zink et al. (2015) and the chapter in this handbook by Antonio Puente and colleagues. Although some interview formats may be less susceptible to diagnostic bias based on race, culturally based diagnostic bias clearly occurs regardless of interview type (Adebimpe, 1981; Bell & Mehta, 1980; Neighbors et al., 1999; Neighbors, Trierweiler, Ford, & Muroff, 2003; Strakowski et al., 2003). There are many examples reported in the literature, but a common finding is that African-Americans are more likely to be diagnosed with schizophrenia (Barnes, 2008; Bresnahan et al., 2007; Minsky, Vega, & Miskimen, 2003; Strakowski et al., 1996; Trierweiler et al., 2000). In some cases, they are three times as likely to be diagnosed with schizophrenia as compared to Caucasians (Bresnahan et al., 2007; Eack, Bahorik, Newhill, Neighbors, & Davis, 2012). The higher rates of schizophrenia diagnoses may be due to cultural distance between client and clinician (e.g., differences in language, values, and expressions of distress), stereotypes of African American psychopathology (e.g., hostility, reluctance to get treatment), false-positive symptoms (e.g., flat affect, paranoia), and biased diagnostic instruments which are not culturally sensitive (Adebimpe, 1981). These and other findings of diagnostic bias are clearly concerning to clinicians because valid and reliable diagnoses are important for prescribing appropriate treatment, identifying individuals at risk for psychiatric disorders, reducing stigmatization, and accessing mental health services, among others (McGuire & Miranda, 2008).

Structured and semistructured interviews might help reduce diagnostic bias because the standardized questions and ordering of information may reduce the possibility of clinician stereotypes that may result in restricting the range of questions. However, to the extent that the interviews were developed on dominant culture populations or employ diagnostic criteria developed on dominant culture, structured interviews in particular may not allow for the flexibility necessary for consideration of culture-based expressions of psychiatric symptoms. Bias may also be introduced while using such measures when client responses are misinterpreted by the

interviewer, thus causing erroneous endorsement of diagnostic criteria. Instruments such as the Composite International Diagnostic Interview (CIDI; [Kessler & Üstün, 2004](#)) and the Mini International Neuropsychiatric Interview (MINI; [Sheehan et al., 1998](#)) were developed for cross-cultural assessment purposes and hold some promise for advancing clinical assessment with ethnic minority populations. As the United States population increases in diversity, clinicians are more likely to evaluate individuals with different cultures from their own, but there is still limited information regarding how various cultures influence the interview process and diagnostic and treatment decision making, and there remain few assessment procedures specifically developed for use with ethnic minorities ([Akin & Turner, 2006](#)).

In the absence of information and procedures, developing cultural sensitivity should be an ongoing goal for clinicians and some have proposed useful conceptual frameworks that can be integrated into the interview process to mitigate diagnostic bias that might arise from culture related factors. The DSM-5 contains sections on cultural formulation and interviews and these materials are made available by the APA online and free of charge at <https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/assessment-measures>. The cultural formulation interview is a semistructured interview composed of 16 questions designed to help the clinician determine the extent to which culture may impact the client's clinical presentation or care ([APA, 2013](#)). It is designed to be used with any individual and employs a person-centered approach to avoid stereotyping.

[Grieger \(2008\)](#) has also developed a conceptual framework that can be integrated into the clinical interview which may prove helpful. The framework, which is grounded in the concepts of world view and acculturation, emphasizes that the following fundamental components should be considered in the clinical interview: (1) problem conceptualization and attitudes towards helping, (2) cultural identity, (3) level of acculturation, (4) family structure and expectations, (5) level of racial/cultural identity development, (6) experiences with bias, (7) immigration issues, (8) existential/spiritual issues, (9) counselor characteristics and behaviors, (10) implications of cultural factors between the counselor and the client, (11) summary of cultural factors and implications for diagnosis, case conceptualization, and treatment. This or another conceptual framework that acknowledges the possible influences of culture on the clinical interview may reduce diagnostic and other bias as culturally sensitive approaches and measures continue to be developed.

## ***Technology***

When the prior edition of this textbook was published in 2000, technology had a major impact on the ways in which individuals communicated professionally and personally through email and computer chats. Use of computers in the healthcare arena allowed for growth of health-care informatics, and video conferencing technology provided a framework for telehealth/telemedicine initiatives to provide mental health services. [Wiens and Brazil \(2000\)](#) concluded "To us, there is little doubt that a transformation in health-care delivery is under way, that computers are the instruments of change, and that communication between patients and medical databases and

between patients and clinicians promises to replace a substantial amount of care now delivered in person" (p. 388). The extent to which technology replaced in-person healthcare delivery is debatable, but there is no doubt that technology has significantly influenced the healthcare delivery system and psychological assessment in expected and unexpected ways. As expected, telehealth continues to grow with the greater availability of video conferencing equipment. Computerized data collection of signs and symptoms is increasingly commonplace. The term "e-health," was coined which is broader than telehealth and includes the use of information and communication technology to connect clinicians with clients in real time across geographical distances to deliver health care immediately (Lal & Adair, 2014). Four areas of mental health service delivery most greatly impacted thus far are (1) information provision, (2) intervention, (3) social support, and (4) screening, assessment, and monitoring. Progress made in screening, assessment and monitoring was to some degree unexpected and fueled by the wide availability of hand-held phones and tablets. These devices allowed for ecological momentary assessment (EMA) of mood and other symptoms (Shiffman, Stone, & Hufford, 2008). EMAs allow real time, remote, and continuous collection of symptom data and address shortcomings of paper and pencil tests such as low compliance and recall bias for retrospective reporting (Shiffman et al., 2008). Initial support exists for the reliability and validity of EMAs for depression, although evidence for mania is limited (Faurholt-Jepsen, Munkholm, Frost, Bardram, & Vedel Kessing, 2016). There are many other relevant technological developments to clinical interviewing, but two that are quite innovative are ongoing efforts to develop virtual patients to train healthcare professionals and the use of computers to conduct clinical interviews. A chapter devoted to technology in this text covers additional information on this topic.

*Virtual patients.* Virtual patients (VPs) are computer based interactive programs that simulate clinical scenarios encountered in clinical practice settings. VPs were developed to teach clinical reasoning and decision making which is a complex skill set that requires integration of professional knowledge with data collected from a client in order to make a diagnosis, prescribe treatment, and monitor client response. VPs vary in style and complexity from rudimentary text-based narratives to sophisticated virtual reality scenarios (Kononowicz, Zary, Edelbring, Corral, & Hege, 2015), the later which have seen broader use due to a dramatic increase in virtual reality technology and a concomitant decrease in technology cost (Parsons & Phillips, 2016). Virtual reality equipment which cost thousands of dollars 10 years ago can currently be purchased for several hundred dollars or less (Parsons et al., 2017). Evidence suggests that use of VPs reinforce students' clinical reasoning abilities (Cook, Erwin, & Triola, 2010), provide a safe environment in which learners can make errors in judgment without adversely effecting real patients (Posel, McGee, & Fleiszer, 2015), allow learners to gain experience with a variety of clinical cases and scenarios, and may be a more cost effective option to teach clinical decision making over other teaching and simulation modalities, such as mannequin-based simulation techniques (Haerling, 2018). Use of VPs also holds promise to provide standardized assessment techniques to evaluate the progress of learners in attaining criterion performance.

[Hege, Kononowicz, Berman, Lenzer, and Kiesewetter \(2018\)](#) note that while clinical reasoning is a fundamental skill it is not fully understood because of its multifactorial nature and as a result, can be difficult to teach. VPs may help in the teaching process, but [Hege et al. \(2018\)](#) suggest that in order for this to occur, five categories should be addressed in VP learning simulations which include “Learner-centeredness (learner), Patient-centeredness (patient), Psychological Theories (researcher), Teaching/Assessment (teacher), and Context (healthcare professional)” (p. 3). Addressing these categories will be an ongoing challenge as new VPs are developed, although VPs are used with increasing frequency in the fields of medicine and nursing to teach a variety of skills including physical examination, medication management, and critical care ([Beal et al., 2017](#); [Tomesko, Touger-Decker, Dreker, Zelig, & Parrott, 2017](#); [Voutilainen, Saaranen, & Sormunen, 2017](#)) with applications to the clinical interviewing. In the coming years, a continued increase is expected in the frequency of VP use to train students in clinical interviewing. Aside from improving general diagnostic skill, VPs may offer a more accessible and effective method to teach structured and semistructured interviewing, thereby increasing the adoption of such assessment procedures by practitioners in clinical settings. VP training may also improve the reliability of flexibly structured interviews which are the most common type of interview. It may also provide a balance between more formally structured and semistructured interviews (which many clinicians perceive as cumbersome) and the completely flexible interview which is prone to unreliability due to criterion and information variance.

*Computer administered clinical interviews and ratings.* A second technological development is the use of computers to conduct clinical interviews. Computer-based interviews are increasingly emerging and like VPs, they range in sophistication. [Reilly-Harrington and coworkers \(2010\)](#) provide an example of one such interview for mania called the Interactive Computer Interview for Mania (ICI-M). The ICI-M is designed to administer the Young Mania Rating scale (YMRS) in a manner like a clinician by using a computer to present scripted probe questions that elicit information about the presence and severity of manic symptoms. Clients respond to consecutive questions in a yes/no or multiple-choice format using a touch screen or mouse click and the next question appears after a response is registered. The ICI flexibly adapts to each client by presenting questions based on the client’s response to the prior question. A proprietary scoring algorithm is used to score each question with the interview and concludes once sufficient information is obtained to complete all YMRS items. Initial reliability and validity studies that compare the ICI-M to the clinician administered YMRS provide support for the scale and overall client satisfaction with the ICI-M was similar to the clinician administered YMRS ([Reilly-Harrington et al., 2010](#)). ICI’s are also available for rating depression and anxiety symptoms. The authors state: “The goal of the ICI is not to replace the human rater, but to provide a standard comparator that can be used to enhance the sensitivity and consistency of human raters...” (p. 522). It remains unknown whether the ultimate outcome of ISIs and other similar assessment procedures will replace the clinician or simply assist the clinician, analogous to computer administration and report generation for certain personality tests.

## **DSM-5**

Development of the DSM criteria provided a common vocabulary for psychiatric disorders that has dramatically influenced clinical interviewing by providing specific criteria sets and organization for the diagnosis of mental disorders. Over the years, the DSM has been revised to reflect new scientific information, with the addition and omission of disorders. Since the last publication of this Handbook, the most recent revision of the DSM was published, the DSM-5 (APA, 2013). It has three main components which include a diagnostic classification system, diagnostic criteria sets, and descriptive text. The diagnostic classification system is where the official DSM disorders are listed along with their diagnostic codes. The diagnostic criteria sets provide detailed criteria for each of the disorder including symptom type and duration required for the diagnosis and other disorders to rule out. The descriptive text provides relevant information for each disorder such as diagnostic features, prevalence, subtypes, risk factors, and culture and gender related diagnostic issues, among others. There is remarkable consistency across revisions since the DSM-III-R, but significant changes also occurred not all of which can be addressed in this chapter. The American Psychiatric Association Website provides extensive information about these changes and interested readers are referred there for additional information <https://www.psychiatry.org/psychiatrists/practice/dsm>.

Directly relevant to the current chapter is the development of assessment measures to assist in the DSM-5 diagnostic process. These assessment measures are available free of charge for use by clinicians and researchers and can be accessed online at <https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources>. The APA refers to these measures as “emerging” because they are under continued development based on feedback from clinicians and researchers, but they are at a stage of development where they may be effectively used in both clinical and research settings. Applications include use in the initial evaluation of the client to help establish and improve reliability of diagnosis. Thereafter, they may be useful to track change in symptoms as a result of intervention and as a means of documenting outcome. They include anchored behavioral ratings, administration, scoring, and interpretation guidelines. Some measures are completed by the clinician based on a clinical interview with the client, and others are self- or informant-report. Five types of measures are provided, including: (1) cross-cutting symptom measures, (2) severity measures that are disorder-specific, (3) The World Health Organization Disability Assessment Schedule, Version 2.0, (4) personality inventories for DSM-5, and (5) early development and home background measures.

The cross-cutting symptom measures are designed to assist in comprehensive mental status assessment of the client. They are separated into two levels; the first level measures provide brief surveys of the major diagnostic domains, while the level 2 measures provide in-depth assessment of specific domains. All measures are completed by the client or an informant and inquire about the presence of symptoms prior to the visit over a specified timeframe, typically one or two weeks. The level 1 measure for adults consists of 23 questions that are designed to assess 13 psychiatric domains such as depression, anger, mania, anxiety, suicidal ideation,

psychosis, and sleep problems, among others. The client rates each item on a 5-point scale of: 0—none or not at all; 1—slight or rare, less than a day or two; 2—mild or several days; 3—moderate or more than half the days; and 4—severe or nearly every day. Based on responses to the level 1 measure, clinicians may opt to provide a more detailed assessment and use the level 2 measures for this purpose. There are separate level 2 measures for 8 of the 13 level 1 domains that inquire about specific symptoms that are used for diagnosis. Self-report and parent-report Level 1 and 2 measure are also available to assess children. These are structured in a similar manner as adult forms, although content varies as developmentally appropriate.

The second type of measures are those based on severity of symptoms for specific disorders. The items in these measures correspond to symptom criteria for the disorder in question. For adults, severity measures are available for assessment of Depression, Separation Anxiety Disorder, Specific Phobia, Social Anxiety Disorder, Panic Disorder, Agoraphobia, Generalized Anxiety Disorder, Posttraumatic Stress Symptoms, Acute Stress Symptoms, and Dissociative Symptoms. The format of these measures varies, as some are self-report and others clinician rated. Timeframes for reporting, item number, and item ratings differ. However, all measures are standardized and come with instructions for administration, scoring and interpretation. For example, depression severity is assessed through self-report using the PROMIS Depression Short Form which consists of 8 items that assess feelings of depression over the past week. Items are rated on a 1–5 scale (1 = Never; 5 = Always) and a total score is derived by summing the item scores. A table is provided that allows conversion of the raw score to a *T* score for interpretation of depression severity (None to slight, Mild, Moderate, Severe). Child severity measures are also available.

The World Health Organization Disability Assessment Schedule, Version 2.0 (WHODAS 2.0; [Ustun et al., 2010](#)) is a 36-item measure designed to assess disability in adults. Disability is assessed across six domains including (1) understanding and communicating; (2) getting around; (3) self-care; (4) getting along with people; (5) life activities; and (6) participation in society. Items reflect specific activities or actions necessary for normal day to day functioning (e.g., remembering to do important things, eating, making new friends). They are rated on a 1–5 point scale that indicates how much difficulty the individual had in completing the activity over the past 30 days (1 = None; 5 = Extreme or cannot do). A total score can be derived by summing the item responses to reflect the degree of functional limitation consistent with the WHO International Classification of Functioning, Disability and Health. Summary scores may also be calculated for each of the six domains. The scale is typically completed by the client, but an informant can complete it when the client is too disabled to do so.

The fourth category of measures are the personality inventories which are designed to measure maladaptive personality traits across five domains: negative affect, detachment, antagonism, disinhibition, and psychotism. Inventories are available for adults and children ages 11–17 and come in two forms, a brief 25-item measure and an extended 220-item measure. The adult extended version may

be completed by an informant if the client is too impaired to complete it. Each item is rated on a 0–3 scale (0 = Very False or Often False; 3 = Very True or Often True) and clients are asked to select “the response that best describes you.” Detailed scoring instructions are provided including procedures to handle missing data. Raw scores can be calculated for each domain, and a total score, and average scores are calculated to allow the clinician to consider severity of the client’s personality dysfunction (mild, moderate, severe) compared to the norm.

The last category consists of two measures designed for assessment of early development and home background. The first measure, the Early Development and Home Background Form—Parent/Guardian (EDHB-PG), is completed by the parent or guardian. It consists of 19 items that assess early development, early communication, and home environment. Examples of item content include premature birth, seizures, hearing and eyesight, hospital admissions, and fights/arguments in the home. Most items are scored “Yes” or “No” with options also included for “Can’t Remember” and “Don’t Know.” The second measure, the Early Development and Home Background Form—Clinician (EDHB-C), is completed by the clinician and consists of 8 items to assess early central nervous system problems, early abuse or neglect, and home environment. According to the instructions, the two measures are used in tandem with the EDHB-PG completed first so that the clinician can review the information. The clinician then conducts a clinical interview with the parent to clarify any information reported on the EDHB-PG. Based on the clinical interview and any other available clinical information, the clinician then completes the EDHB-C.

The impact of these procedures on the clinical diagnostic process is yet to be determined. Their availability and continued development have the potential to improve the reliability of diagnoses by introducing standardization and sequencing into the assessment process. They can be selected based on clinical need and in this sense are possibly more “clinician friendly” than other more extensive structured and semistructured procedures used in research settings. The fact that they were included in the DSM-5 acknowledges the importance of using standardized procedures to make psychiatric diagnosis, which falls in line with many years of research advocating such procedures. Easily implemented standardized assessment procedures in practice settings allow more efficient means of providing clinical services in the context of the managed care environment and will undoubtedly contribute to whether or not clinicians utilize them in clinical practice.

## Conclusion

Clinical interviewing has dramatically evolved over the past century. Development of widely accepted criteria for diagnosis advanced efforts to standardize and sequence the interview process for diagnostic purposes. Changes to healthcare delivery systems and the managed care environment imposed external constraints on the interview, requiring that detailed and often extensive information is collected

in a brief period. Contemporaneously, awareness grew regarding the need for improving reliability of symptom and behavioral data by minimizing criterion and information variance. Researchers and clinicians addressed this by developing many standardized assessment procedures to assist with diagnosis and documenting presence and severity of psychopathological symptoms. The recent publication of the next iteration of the DSM includes some of these assessment procedures but their utility in clinical settings awaits further development and feedback. Technology increasingly influenced the provision of healthcare services and influenced clinical interviewing in ways that could not be anticipated since the last edition of this Handbook. However, this influence will likely increase in subsequent years in ways that are both predictable and unpredictable. Whatever form the clinical interview takes in the future, it will continue to hold an important place in the provision of mental and medical health care.

## Acknowledgments

This work represents a collection of efforts from many people. The authors would like to extend special thanks to Jayson Wright, Anita Kwong, and Sarah Flood for their contributions to this chapter.

## References

- Adebimpe, V. R. (1981). Overview: White norms and psychiatric diagnosis of Black patients. *American Journal of Psychiatry, 138*(3), 279–285.
- Adebimpe, V. R. (1994). Race, racism, and epidemiological surveys. *Hospital and Community Psychiatry, 45*, 27–31.
- Akin, W. M., & Turner, S. M. (2006). Toward understanding ethnic and cultural factors in the interviewing process. *Psychotherapy: Theory, Research, Practice, Training, 43*(1), 50–64.
- Al-issa, I. (1995). The illusion of reality or the reality of illusion: Hallucinations and culture. *British Journal of Psychiatry, 166*, 368–373.
- American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed). Arlington, VA: American Psychiatric Publishing.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed). Arlington, VA: American Psychiatric Publishing.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed). Arlington, VA: American Psychiatric Publishing.
- Barnes, A. (2008). Race and hospital diagnoses of schizophrenia and mood disorders. *Social Work, 53*, 77–83.
- Beal, M. D., Kinnear, J., Anderson, C. R., Martin, T. D., Wamboldt, R., & Hooper, L. (2017). The effectiveness of medical simulation in teaching medical students critical care medicine a systematic review and meta-analysis. *Simulation in Healthcare, 12*(2), 104–116. Available from <https://doi.org/10.1097/SIH.0000000000000189>.

- Bell, C., & Mehta, H. (1980). The misdiagnosis of black patients with manic depressive illness. *Journal of the National Medical Association*, 72, 141–145.
- Benjamin, A. (1969). *The helping interview*. Boston: Houghton-Mifflin.
- Blashfield, R. K. (1992). *Are there any prototypical patients with personality disorders?* Paper presented at the American Psychological Association convention, Washington, DC.
- Bresnahan, M., Begg, M. D., Brown, A., Schaefer, C., Sohler, N., Insel, B., . . . Susser, E. (2007). Race and risk of schizophrenia in a US birth cohort: Another example of health disparity? *International Journal of Epidemiology*, 36, 751–758.
- Cook, D. A., Erwin, P. J., & Triola, M. M. (2010). Computerized virtual patients in health professions education: A systematic review and meta-analysis. *Academic Medicine*, 85 (10), 1589–1602. Available from <https://doi.org/10.1097/ACM.0b013e3181edfe13>.
- Dana, R. H. (2008). Clinical diagnosis in multicultural populations. In L. A. Suzuki, & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp. 107–131). San Francisco, CA: John Wiley & Sons.
- Deutsch, F., & Murphy, W. F. (1955a). *The clinical interview: Vol. 1. Diagnosis*. New York: International Universities Press.
- Deutsch, F., & Murphy, W. F. (1955b). *The clinical interview: Vol. 2. Therapy*. New York: International Universities Press.
- Donnelly, J., Rosenberg, M., & Fleeson, W. P. (1970). The evolution of the mental status—Past and future. *American Journal of Psychiatry*, 126, 997–1002.
- Eack, S. M., Bahorik, A. L., Newhill, C. E., Neighbors, H. W., & Davis, L. E. (2012). Interviewer-perceived honesty as a mediator of racial disparities in the diagnosis of schizophrenia. *Psychiatric Services*, 63(9), 875–880.
- Edelbrock, C., & Costello, A. J. (1984). Structured psychiatric interviews for children and adolescents. In G. Goldstein, & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 276–290). New York: Pergamon Press.
- Edelbrock, C., & Bohnert, A. (2000). Structured interviews for children and adolescents. In G. Goldstein, & M. Hersen (Eds.), *Handbook of psychological assessment* (3rd ed, pp. 369–386). New York: Pergamon.
- Egan, G. (1975). *The skilled helper: A model for systematic helping and interpersonal relating*. Belmont, CA: Brooks/Cole.
- Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The schedule for affective disorders and schizophrenia. *Archives of General Psychiatry*, 35, 837–844.
- Faurholt-Jepsen, M., Munkholm, K., Frost, M., Bardram, J. E., & Vedel Kessing, L. (2016). Electronic self-monitoring of mood using IT platforms in adult patients with bipolar disorder: A systematic review of the validity and evidence. *BMC Psychiatry*, 16(7). Available from <https://doi.org/10.1186/s12888-016-0713-0>.
- Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry*, 26, 57–63.
- Freud, A. (1936). *The ego and the mechanisms of defense*. New York: International Universities Press.
- Gill, M., Newman, R., & Redlich, F. C. (1954). *The initial interview in psychiatric practice*. New York: International Universities Press.
- 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* (pp. 132–161). San Francisco, CA: John Wiley & Sons.
- Grinder, J., & Bandler, R. (1975). *The structure of magic*. Palo Alto: Science & Behavior.

- Haerling, K. A. (2018). Cost-utility analysis of virtual and mannequin-based simulation. *Simulation in Healthcare*, 13(1), 33–40. Available from <https://doi.org/10.1097/SIH.0000000000000280>.
- Hall, C. S., & Lindzey, G. (1978). *Theories of personality*. New York: Wiley.
- Hartmann, H. (1939). *Ego psychology and the problem of adaptation*. (p. 1958) New York: International Universities Press.
- Havens, L. (1978). Explorations in the uses of language in psychotherapy: Simple empathic statements. *Psychiatry*, 41, 336–345.
- Havens, L. (1979). Explorations in the uses of language in psychotherapy: Complex empathic statements. *Psychiatry*, 42, 40–48.
- Hege, I., Kononowicz, A. A., Berman, N. B., Lenzer, B., & Kiesewetter, J. (2018). Advancing clinical reasoning in virtual patients—development and application of a conceptual framework. *GMS Journal for medical education*, 35(1), Doc12.
- Henderson, M. C., Tierney, L. M., Jr., & Smetana, G. W. (2012). *The patient history: An evidence-based approach to differential diagnosis* (2nd ed). China: McGraw-Hill.
- Hersen, M., & Sturmey, P. (2012). *Handbook of evidence-based practice in clinical psychology, adult disorders* (Vol. 2). John Wiley & Sons.
- Hersen, M., & Turner, S. M. (1985). *Diagnostic interviewing*. New York: Plenum Press.
- Kaplan, H., Freedman, A., & Sadock, B. (Eds.), (1980). *Comprehensive textbook of psychiatry* (3rd ed). Baltimore: Williams & Wilkins.
- Kessler, R. C., & Üstün, T. B. (2004). The world mental health (WMH) survey initiative version of the world health organization (WHO) composite international diagnostic interview (CIDI). *International Journal of Methods in Psychiatric Research*, 13(2), 93–121.
- Kononowicz, A. A., Zary, N., Edelbring, S., Corral, J., & Hege, I. (2015). Virtual patients—what are we talking about? A framework to classify the meanings of the term in health-care education. *BMC Medical Education*, 15(11). Available from <https://doi.org/10.1186/s12909-015-0296-3>.
- Lal, S., & Adair, C. E. (2014). E-mental health: A rapid review of the literature. *Psychiatric Services*, 65(1), 24–32. Available from <https://doi.org/10.1176/appi.ps.201300009>.
- Langsley, D. C., & Yager, J. (1988). The definition of a psychiatrist: Eight years later. *American Journal of Psychiatry*, 145, 469–475.
- Lecrubier, Y., Weiller, E., Hergueta, T., Amorim, P., Bonora, L. I., Lépine, J. P., Sheehan, D., Janavs, J., Baker R., Sheehan, K. H. (1998). *M.I.N.I. Mini International Neuropsychiatric Interview, French Version, 5.0.0., Vie entière, DSM-IV*. DOI: 10.13140/RG.2.1.2792.9440.
- Lewis-Fernández, R., Hinton, D. E., Laria, A. J., Patterson, E. H., Hofmann, S. G., Craske, M. G., ... Liao, B. (2010). Culture and the anxiety disorders: Recommendations for DSM-V. *Depression and Anxiety*, 27(2), 212–229.
- MacKinnon, R. A., & Michels, R. (1971). *The psychiatric interview in clinical practice*. Philadelphia: W.B. Saunders.
- Margulies, A. (1984). Toward empathy: The uses of wonder. *American Journal of Psychiatry*, 141, 1025–1033.
- Margulies, A., & Havens, L. (1981). The initial encounter: What to do first? *American Journal of Psychiatry*, 138, 421–428.
- McGuire, T. G., & Miranda, J. (2008). New evidence regarding racial and ethnic disparities in mental health: Policy implications. *Health Affairs*, 27, 393–403.
- Meyer, A. (1951). In E. E. Winters (Ed.), *The collected papers of Adolf Meyer* (Vol. 3). Baltimore: John Hopkins Press.

- Minsky, S., Vega, W., Miskimen, T., et al. (2003). Diagnostic patterns in Latino, African American, and European American psychiatric patients. *Archives of General Psychiatry*, 60, 637–644.
- Neighbors, H. W., Jackson, J. S., Campbell, L., & Williams, D. (1989). The influence of racial factors on psychiatric diagnosis: A review and suggestions for research. *Community Mental Health Journal*, 25, 301–309.
- Neighbors, H. W., Trierweiler, S. J., Ford, B., & Muroff, J. R. (2003). Racial differences in diagnosis using a semi-structured instrument: The importance of clinical judgment. *Journal of Health and Social Behavior*, 44, 237–256.
- Neighbors, H. W., Trierweiler, S. J., Munday, C., Thompson, E. E., Jackson, J. S., Binion, V. J., & Gomez, J. (1999). Psychiatric diagnosis of African Americans: Diagnostic divergence in clinician-structured and semi-structured interviewing conditions. *Journal of the National Medical Association*, 91(11), 601–612.
- Othmer, E., & Othmer, S. C. (1989). *The clinical interview using DSM-III-R*. Washington, DC: American Psychiatric Press, Inc.
- Othmer, E., & Othmer, S. C. (1994a). *The clinical interview using DSM-IV: Vol. 1. Fundamentals*. Washington, DC: American Psychiatric Press, Inc.
- Othmer, E., & Othmer, S. C. (1994b). *The clinical interview using the DSM-IV: Vol. 2. The difficult patient*. Washington, DC: American Psychiatric Press, Inc.
- Paniagua, F. A. (2001). Culture-bound syndromes, cultural variations, and psychopathology. In I. Cuellar, & F. A. Paniagua (Eds.), *Handbook of multicultural mental health* (pp. 139–169). San Diego, CA: Academic Press.
- Parsons, T. D., & Phillips, A. (2016). Virtual reality for psychological assessment in clinical practice. *Practice Innovations*, 1, 197–217.
- Parsons, T. D., Riva, G., Parsons, S., Mantovani, F., Newbutt, N., Lin, L., ... Hall, T. (2017). Virtual reality in pediatric psychology: Benefits, challenges, and future directions. *Pediatrics*, 140, 86–91.
- Pascal, G. R. (1983). *The practical art of diagnostic interviewing*. Homewood, IL: Dow Jones-Irwin.
- Pomales, J., Claiborn, C. D., & Lafromoise, T. D. (1986). Effects of black-students racial identity on perceptions of white counselors varying in cultural sensitivity. *Journal of counseling psychology*, 33(1), 57–61.
- Posel, N., McGee, J. B., & Fleiszer, D. M. (2015). Twelve tips to support the development of clinical reasoning skills using virtual patient cases. *Medical Teaching*, 37(9), 813–818. Available from <https://doi.org/10.3109/0142159X.2014.993951>.
- Reik, T. (1952). *Listening with the third ear*. New York: Farrar, Strauss.
- Reilly-Harrington, N. A., DeBonis, D., Leon, A. C., Sylvia, L., Perlis, R., Lewis, D., & Sachsa, G. S. (2010). The interactive computer interview for mania. *Bipolar Disorders*, 12(5), 521–527. Available from <https://doi.org/10.1111/j.1399-5618.2010.00844.x>.
- Richardson, S. A., Dohrenwend, B. S., & Klein, D. (1965). *Interviewing: Its forms and functions*. New York: Basic Books.
- Robins, L. N., Cottler, L. B., Bucholz, K., & Compton, W. (1995). *NIMH diagnostic interview schedule, version IV*. Washington School of Medicine: St. Louis.
- 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*, 10, 41–61.
- Rogers, C. R. (1951). *Client-centered therapy*. Boston: Houghton-Mifflin.
- Rogers, C. R. (1959). A theory of therapy, personality and interpersonal relationships as developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A study of a*

- science: *Formulations of the person and the social context* (Vol. 3). New York: McGraw-Hill.
- Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. New York, NY: Guilford Press.
- Rollnick, S., & Miller, W. R. (1995). What is motivational interviewing? *Behavioural and Cognitive Psychotherapy*, 23(4), 325–334.
- Saywitz, K. J., & Campano, L. B. (2013). *Evidence-based child forensic interviewing: The developmental narrative elaboration interview*. Oxford University Press.
- Shea, S. C. (1988). *Psychiatric interviewing: The art of understanding*. Philadelphia: W.B. Saunders.
- Shea, S. C. (1988a). *Psychiatric interviewing: The art of understanding*. Philadelphia: W.B. Saunders.
- Shea, S. C. (1998b). The chronological assessment of suicide events: A practical interviewing strategy for the elicitation of suicidal ideation. *Journal of Clinical Psychiatry*, 59 (Suppl), 58–72.
- Shea, S. C. (1999). *The practical art of suicide assessment*. New York: John Wiley & Sons, Inc.
- Shea, S. C. (2000). Contemporary clinical interviewing: Integration of the DSM-IV, managed care concerns, mental status, and research. In G. Goldstein, & M. Hersen (Eds.), *Handbook of psychological assessment* (3rd ed, pp. 339–368). New York: Pergamon.
- Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., ... Dunbar, G. C. (1998). The mini-international neuropsychiatric interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, 59, 22–33.
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32.
- Spitzer, R. G., Endicott, J., & Robins, E. (1975). Clinical criteria for diagnosis and DSM-III. *American Journal of Psychiatry*, 132, 1187–1192.
- Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria. *Archives of General Psychiatry*, 35, 773–782.
- Spitzer, R. L., & Williams, J. B. W. (1983). *DSM-III SCID manual*. New York: New York State Psychiatric Institute, Biometrics Research Department.
- Strakowski, S. M., Flaum, M., Amador, X., Bracha, H. S., Pandurangi, A. K., Robinson, D., & Tohen, M. (1996). Racial differences in the diagnosis of psychosis. *Schizophrenia Research*, 21(2), 117–124.
- Strakowski, S. M., Keck, P. E., Jr., Arnold, L. M., Collins, J., Wilson, R. M., Fleck, D. E., et al. (2003). Ethnicity and diagnosis in patients with affective disorders. *Journal of Clinical Psychiatry*, 64(7), 747–754.
- Sullivan, H. S. (1970). *The psychiatric interview*. New York: W. W. Norton Company.
- Tomesko, J., Touger-Decker, R., Dreker, M., Zelig, R., & Parrott, J. S. (2017). The effectiveness of computer-assisted instruction to teach physical examination to students and trainees in the health sciences professions: A systematic review and meta-analysis. *Journal of Medical Education and Curricular Development*, 4. Available from <https://doi.org/10.1177/2382120517720428>.
- Trierweiler, S. J., Neighbors, H. W., Munday, C., Thompson, E. E., & Binion, V. J. (2000). Clinician attributions associated with the diagnosis of schizophrenia in African American and non-African-American patients. *Journal of Consulting and Clinical Psychology*, 68, 171–175.

- Ustun, T. B., Chatterji, S., Kostanjsek, N., Rehm, J., Kennedy, C., Epping-Jordan, J., ... in collaboration with WHO/NIH Joint Project. (2010). Developing the world health organization disability assessment schedule 2.0. *Bulletin of the World Health Organization*, 88, 815–823.
- Voutilainen, A., Saaranen, T., & Sormunen, M. (2017). Conventional vs. e-learning in nursing education: A systematic review and meta-analysis. *Nurse Education Today*, 50, 97–103.
- Want, V., Parham, T. A., Baker, R. C., & Sherman, M. (2004). African American students' ratings of Caucasian and African American counselors varying in racial consciousness. *Cultural Diversity and Ethnic Minority Psychology*, 10, 123–136.
- Ward, C. H., Beck, A. T., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1962). The psychiatric nomenclature. *Archives of General Psychiatry*, 7, 198–205.
- Whitehorn, J. C. (1944). Guide to interviewing and clinical personality study. *Archives of Neurology and Psychiatry*, 52, 197–216.
- Wiens, A. N., & Brazil, P. J. (2000). Structured clinical interviews for adults. In G. Goldstein, & M. Hersen (Eds.), *Handbook of psychological assessment* (3rd ed, pp. 387–412). New York: Pergamon.
- Zink, D., Lee, B., & Allen, D. N. (2015). Structured and semi-structured clinical interviews available for use among African American clients: Cultural considerations in the diagnostic interview process. In L. Benuto, & B. D. Leany (Eds.), *Guide to psychological assessment with African-Americans* (pp. 19–42). New York, NY: Springer Publishing, ISBN 978-1-4939-1003-8.

# Structured and semistructured interviews for children

11

Christopher A. Kearney, Andrew Freeman and Victoria Bacon

Department of Psychology, University of Nevada, Las Vegas, NV, United States

## Introduction

Clinicians may choose from a variety of assessment tools for child clinical assessment (e.g., questionnaires, intelligence tests, neuropsychological tests) depending on both their training and preferences as well as client characteristics such as age and presenting question. However, almost all clinicians conduct clinical interviews as part of the assessment process (Shapiro & Heick, 2004; Watkins, Campbell, Nieberding, & Hallmark, 1995). Therefore, the clinical interview is the hallmark of clinical assessment.

Many definitions of a clinical interview have been offered. A clinical interview can generally be defined as a purposeful interaction focused on gathering information regarding a child's behavioral, social, and emotional functioning in regard to planning, implementing, or evaluating an intervention (McConaughy, 2013; Merrell, 2008). Most broadly speaking, a child clinical interview consists of a formally arranged meeting with a specific purpose in which there is a defined relationship between the interviewer and interviewee. Common interviewees are the child, caregiver, and teacher (Sattler, 2008). The interviewer uses purposeful questioning to direct the flow of the conversation through choosing questions, topics, or the broad content of the discussion. The interviewer attends nonjudgmentally to both the content and other aspects of the interaction (e.g., affect, behavior, style). Finally, the interviewer follows guidelines regarding confidentiality of the information gathered. However, these are broad definitions that encompass many types of interviews that are used for both research and clinical practice. Child clinical interviews in practice typically focus on a child's strengths (e.g., individual, family relationships, resources, goals, and aspirations), weaknesses (e.g., economic, discrimination, individual vulnerabilities, family vulnerabilities), culture (e.g., expectations and rituals), and desired outcomes (e.g., child and family inclusion in treatment planning).

## Historical perspectives

At the turn of the 20th century, clinical assessment was strongly informed by early cognitive tests, observation, and the beginnings of psychoanalysis. For example,

Emil Kraepelin relied heavily on longitudinal observations of patients' behavior to define mood disorders and psychosis (Kraepelin, 1921). In contrast to many contemporaries who focused on developing cognitive ability tests that could be used to place individuals relative to each other, Jean Piaget developed the semiclinical interview for the purpose of probing a child's thinking (Elkind, 1964). In the late 1920s, Piaget developed a series of interviews to examine children's understanding of causality, the world and judgment. These early interviews combined both standardized, structured inquiries and open-ended questions. In modern terminology, Piaget's early interviews screened a child's cognitive ability within his developmental staging framework. Modern interviews are typically used to provide initial clinical assessments of strengths and weaknesses, make psychiatric diagnoses, inform intervention, evaluate the effectiveness of current services, and screen for at-risk status.

Child clinical interviewing reemerged in the 1950s and 1960s in contrast to the general practice of psychodynamic and behavioral assessment strategies. The first two child clinical interviews were focused on examining the prevalence of problematic behaviors in children. In the 1965, Lapouse and Monk developed a structured interview consisting of 200 questions that required mostly yes and no responses from the interviewee. From the community, children ages 6–12 years ( $n = 482$ ) were recruited. Both the children and their mothers were interviewed regarding behavioral problems. Child- and parent-report tended to agree for observable behaviors (e.g., bedwetting); however noticeable differences did occur. Parent-report and child-report were substantially discrepant for fear and worries about the child's safety (Lapouse & Monk, 1959). Additionally, parent-report demonstrated substantial test-retest reliability.

In the mid-1960s, Rutter and Graham developed a descriptive clinical interview focused on child's emotional, behavioral, and social functioning as part of the Isle of Wight studies. Rutter and Graham's interview was semistructured. The parent and child versions were parallel but different (Graham & Rutter, 1968; Rutter & Graham, 1968). Interviewers were provided with functional domains to assess (e.g., school performance, activities, and friendships). Interviewers were allowed to query about the functional domains using their own words and in any order. If psychiatric concerns were mentioned during this section of the interview, the interviewer queried using 36 specific questions. Parents tended to report more detail in regards to duration, severity, action taken, presumed cause, and expected course than children. Similar to Lapouse and Monk, Rutter and Graham demonstrated poorer agreement between child- and parent-report for internalizing difficulties (e.g., depression and anxiety) and better agreement for observable difficulties (e.g., attention, hyperactivity, and antisocial behavior). Most significantly cross-informant agreement was highest for global psychiatric status.

Clinical child interviews would change with the introduction of DSM-III in 1980. DSM-III focused on improving the reliability of psychiatric diagnoses by reducing the focus on psychodynamic prototypes and increasing the focus on observable symptoms. For children, the DSM-III represented a rapid expansion in the number of diagnoses available from two—Adjustment Reaction of Childhood

and Childhood Schizophrenia—to most of the diagnoses available to adults. Many of these children were previously undiagnosed or classified as having Adjustment Reaction (Rosen, Bahn, & Kramer, 1964). DSM-III introduced symptom lists for disorders that were made into diagnostic interviews for adults to improve the reliability of diagnoses for research. Two early interviews for adults—the Diagnostic Interview Schedule (Robins, Helzer, Croughan, & Ratcliff, 1981) and the Schedule for Affective Disorders (Endicott & Spitzer, 1978)—inspired the development of similar interviews for children—the Diagnostic Interview for Children and Adolescents (Herjanic & Reich, 1982) and the Schedule for Affective Disorders and Schizophrenia for School-Age Children (Chambers et al., 1985). These child clinical interviews represent the diagnostic criteria for DSM-oriented disorders in structured and semistructured formats. Revised versions of these interviews continue to be commonly used to this day and the purpose of these diagnostic interviews remains to be increased reliability of the diagnosis across clinicians.

Modern clinical child interviews continue both the descriptive and diagnostic traditions of the past. For example, at intake many clinicians use relatively unstandardized clinical interviews based on life domains (Jones, 2010). These interviews often query about a child's strengths (e.g., individual, family relationships, resources, goals, and aspirations), weaknesses (e.g., economic, discrimination, individual vulnerabilities, family vulnerabilities), culture (e.g., expectations and rituals), and desired outcomes (e.g., child and family inclusion in treatment planning). This tradition of child clinical interviewing continues the early descriptive interviews. In contrast to the unstructured, descriptive tradition, modern diagnostic interviews for children tend to be structured or semistructured. Modern diagnostic interviews encourage information gathering from multiple informants. The two most dominant trends in development of child clinical interviews for diagnosis are: expanded age ranges (e.g., preschool interviews; Egger et al., 2006) and specialized interviews for specific disorders (e.g., anxiety disorders; Silverman & Nelles, 1988).

Structured or semistructured interviews regarding child and adolescent psychiatric disorders generally include two main types. The first type involves general or overarching interviews that cover a wide range of mental health conditions, including emotional, behavioral, substance use, developmental, and psychotic disorders. These interviews are often, though not always, aligned closely with the DSM diagnostic system and its various editions. The second type involves specific interviews that cover one or a few psychiatric conditions in much greater depth. The following sections cover both sets of structured and semistructured interviews.

## General structured interviews

General structured or semistructured interviews include legacy instruments such as the Schedule for Affective Disorders and Schizophrenia for School-Aged Children, Diagnostic Interview for Children and Adolescents, and Diagnostic Interview for Children. These have been widely used in clinical research settings to examine

epidemiological patterns of psychiatric disorders as well as to serve as pre–post measures of treatment outcome. The interviews are typically modified with changes in the DSM categorical system, though long lag times following each new edition are common. DSM cutoffs for clinical determination are commonly adopted by these interviews. The interviews are less commonly used in clinical settings for individual patients, in part because of their length and detail, but may be used as screening devices in some such settings. The following sections cover the most common general interviews.

### ***Schedule for Affective Disorders and Schizophrenia for School-Aged Children***

One of the most venerable semistructured interviews for children and adolescents is the *Schedule for Affective Disorders and Schizophrenia for School-Aged Children: Present and Lifetime Version* (K-SADS-PL) (Ambrosini, 2000). The interview is primarily designed for youth aged 7–18 years and is used to assess a wide range of current and lifetime psychiatric disorders. The interview is utilized with parents and youth, and an interviewer coalesces data from each informant to derive diagnostic assignments via clinical judgment. Administration time may last from 35 to 80 min per informant depending on the severity and complexity of the presenting problems.

The K-SADS-PL includes an introductory interview about general topics related to demographics, current symptoms, past psychiatric intervention, and school, peer, and family functioning. The core of the K-SADS-PL, however, is the screen interview that focuses on 82 symptoms across various psychiatric diagnoses. Ratings of the items are typically made on a 0–3 scale in increasing level of symptom severity (e.g., a score of “3” indicates a fully clinical or “threshold” symptom). Ideally, these symptoms are rated when a particular youth is not on medication, or prior to starting a new medication, to identify the true nature of the symptoms (Kaufman & Schweder, 2004). Raters may add notes regarding their behavioral observations of the child, such as his or her social withdrawal, anxiety, or noncompliance. Positive indication of threshold symptoms generally leads to diagnostic supplements related to mood, anxiety, psychotic, and behavioral disorders as well as substance use, eating, and tic disorders.

Lauth and colleagues (2010) conducted a detailed psychometric analysis of the K-SADS-PL among 86 youth with severe emotional or behavioral difficulties, most typically major depression as well as anxiety, attention-deficit hyperactivity, oppositional defiant, and conduct disorders. Interrater reliability of the interview was generally satisfactory ( $\kappa$ , 0.44–1.00) except for the mania diagnosis (0.31). Strongest psychometric support was found for the depressive disorders, which evinced a 100% sensitivity rate and a 35% specificity rate, with a true positive rate of 86.3%. Other diagnoses generally fared less well, particularly in cases involving substantial comorbidity as well as attention-deficit/hyperactivity disorder. In addition, correlations of diagnoses with specific rating scales of various conditions such

as depression or disruptive behavior were generally moderate, depending on the specificity of the scale.

Different historical versions of this interview are available; the most common include versions for epidemiological data (K-SADS-E) and present state only (K-SADS-P). These versions have variable data sets and results with respect to psychometric support. The K-SADS-P, for example, has generally enjoyed stronger interrater and test–retest reliability across diagnostic categories than the K-SADS-E (Ambrosini, 2000). These historical interview versions tend to be much broader in scope across symptoms and diagnoses than the K-SADS-PL, which may contribute to their more moderate psychometric strength.

### ***Diagnostic Interview for Children and Adolescents***

Another venerable semistructured interview for children and adolescents is the *Diagnostic Interview for Children and Adolescents* (DICA) (Reich, 2000). This interview is designed for 6- to 17-year-old youth and focuses specifically on this informant group. The interview covers over 20 general diagnostic conditions such as attention-deficit/hyperactivity disorder, anxiety and mood disorders, various substance use disorder categories, eating and elimination disorders, and oppositional defiant and conduct disorders. Other general questions cover psychotic symptoms, perinatal factors, psychosocial functioning, impairment, and risk and protective factors (Rourke & Reich, 2004).

Diagnostic sections begin with rapport-building inquiries and then specific yes–no questions about symptoms and the emotional states surrounding these symptoms. As with most general semistructured interviews, the DICA is scripted and the interviewer is instructed to move to a different section depending on the responses provided. Interviewers are, however, encouraged to provide contextual information throughout the interview and note items that may require more detailed follow-up. Diagnostic assignments are manual-based. Less psychometric information is available on the DICA compared to other semistructured interviews, though test–retest reliability tends to be moderate for most diagnostic categories and better for depression ( $\kappa$ , 0.80) and conduct disorder ( $\kappa$ , 0.92) (Reich, 2000).

### ***Diagnostic Interview for Children***

Another legacy semistructured interview is the *Diagnostic Interview for Children-version IV* (DISC-IV; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000). This interview is designed for 6- to 17-year-old youth and has versions for parents and youth. The interview covers over 30 of the most common psychiatric disorders in youth and is quite extensive across anxiety, mood, psychotic, disruptive behavior, substance use, and other psychiatric disorders (nearly 2500 questions total). The interview is largely scripted and thus not designed to rely heavily on clinical judgment. Most questions are in yes–no format, though some items are open-ended in nature. Administration time may last from 45 to 120 min per informant depending on the severity of the presenting problems.

Initial questions in the DISC surround demographic information, age of onset of symptoms, impairment, memorable events, and past intervention. A core or stem set of questions (358) are then asked to provide broad indicators of problems, and are followed by nearly 1300 contingent questions based on previous answers. Contingent questions include information on frequency, duration, and intensity, and several transdiagnostic constructs such as irritability are covered in multiple sections. The DISC-IV examines symptomatology and diagnoses in the past month and past year. A concentration is made on anxiety, mood, disruptive, substance use, and miscellaneous disorders as well as schizophrenia (Shaffer, Fisher, & Lucas, 2004). Whole-life questions (499) are asked in cases of subthreshold or absent symptoms and focus on whether some diagnoses may have applied beyond the past year when the child was younger. Past episodes of psychopathology after age 5 years may thus be documented.

The interview is typically computer-scored across 34 psychiatric diagnoses (Fisher, Lucas, Lucas, Sarsfield, & Shaffer, 2006). Test-retest reliability of the DISC-IV is best for disorders such as specific phobia (0.86) but generally moderate across other diagnoses. Reliability for the parent version tends to be better than the youth version. Other versions of the DISC include the DISC-PS (present state only), DISC-T (for teachers), and DISC Predictive Scales (for predicting disorders on a later DISC administration) as well as versions that are self-administered, applicable to young adults, and shorter in length (Shaffer et al., 2000).

### ***Child and Adolescent Psychiatric Instrument***

The *Child and Adolescent Psychiatric Instrument* (CAPA) is a semistructured interview that focuses on youth aged 9–18 years and has child and parent versions (Angold et al., 1995). This interview includes an introductory section focusing on general functioning with respect to school, home, family life, peers, and activities. This is followed by a symptom review section that covers major diagnostic categories related to anxiety and mood disorders, substance use, food-related and sleep problems, disruptive behavior disorders, somatization disorder, and tic disorders and trichotillomania. Attention-deficit/hyperactivity disorder is covered in the parent version only. The last section, an incapacity section, focuses on psychosocial impairment for each of the symptom domains.

Questions are variable in format, can include information regarding context, and concentrate on functioning during the past 3 months. Administration time is approximately 90 min. Questions include screening items that determine whether or not to pursue a particular section in more depth, emphasized probes or required questions in a section once a screening question has been endorsed, and discretionary probes or suggested additional questions that may provide greater context and clarification regarding the symptoms. Open-ended questions are also allowed, and observations of interview behavior are recorded. Diagnoses are based on a glossary method that includes formal definitions of symptoms, and intensity ratings are assigned as well. Symptom intensity is rated according to its match to the glossary definition.

The psychometric strength of the CAPA is generally in line with most large-scale interviews, with  $\kappa$  values in the moderate range.  $\kappa$  Values tend to be better for depression (0.90) as well as posttraumatic disorder and substance use (0.90+). Diagnostic ratios generally match expected ratios across gender and age groups and utilization of mental health services. Other versions of the CAPA focus on twins, young adults, preschoolers, Spanish speakers, and shorter length ([Angold & Costello, 2000](#)). As with other large-scale interviews, the CAPA has been used primarily in large-scale community and epidemiological studies.

### ***Children's Interview for Psychiatric Syndromes***

The *Children's Interview for Psychiatric Syndromes* (ChIPS) is a structured diagnostic interview for youth aged 6–18 years ([Weller, Weller, Fristad, & Rooney, 1999](#)). The ChIPS is more structured than other interviews described in this chapter, with specific screening or “cardinal” questions that must be endorsed prior to proceeding to a specific section. Cardinal questions involve those symptoms most commonly endorsed by children with a particular psychiatric disorder. Child and parent versions are available. Administration time per informant is approximately 20–50 min; the interview is briefer than ones previously covered in this chapter.

The ChIPS covers major diagnostic categories in addition to several that are not always assessed in other measures, such as acute stress disorder, obsessive-compulsive disorder, hypomania, and encopresis. Furthermore, the ChIPS evaluates different forms of certain disorders, including attention-deficit/hyperactivity disorder (inattentive, hyperactive, combined), nocturnal and diurnal enuresis, conduct disorder (mild, moderate, severe), and obsessive-compulsive disorder (obsessions, compulsions, both). The interview also covers psychosocial stressors and child maltreatment history.

Several studies have examined the psychometric strength of the ChIPS. Diagnoses derived from the ChIPS generally match diagnoses from other interviews, with moderate  $\kappa$  values, among inpatient and outpatient samples. In addition, moderate child-parent agreement ( $\kappa$  value, 0.41) has been noted. Sensitivity has been noted at 87%, with 76% specificity. The psychometric strength of the measure tends to be better among girls, adolescents, and nonclinical populations ([Weller, Weller, Fristad, Rooney, & Schechter, 2000](#)).

### ***Mini International Neuropsychiatric Interview for Children and Adolescents***

The *Mini International Neuropsychiatric Interview for Children and Adolescents* (MINI-KID) was initially designed as a bridge between lengthy structured diagnostic interviews and very brief rating scales used as screening devices. The child version is based on the original MINI ([Sheehan et al., 1998](#)). The MINI-KID is designed for youth aged 6–17 years and is organized along diagnostic modules,

each of which begins with 2–4 screening questions. Child and parent versions are available, and the interview is designed to be administered in 30 min.

The MINI-KID is unique in that various diagnostic categories are examined across different timelines. Some diagnoses are evaluated for symptomatology in a very recent timespan, such as 2 weeks, as is the case for depression and adjustment disorders. Other diagnoses are assessed with respect to a 1-month timespan, most notably the anxiety disorders but also suicidality and tic disorders. Some diagnoses are assessed utilizing a 3-month timespan, including eating disorders, and other diagnoses are assessed utilizing a 6-month timespan, such as generalized anxiety, oppositional defiant, and attention-deficit/hyperactivity disorders. Longer timeframes up to 1 year are utilized for alcohol and substance use problems, conduct disorder, and dysthymia. Current symptomatology for all disorders is assessed as well, including pervasive developmental disorder.

The psychometric strength of the MINI-KID tends to resemble other structured interviews, though the measure also tends to produce more overall diagnoses.  $\kappa$  Values for major diagnostic categories are in the 0.41–0.87 range, higher for substance use disorder and lower for psychotic disorder. Interrater reliability tends to be strong for broad syndromes (0.94) and individual diagnoses (0.89); test–retest reliability is also favorable for broad syndromes (0.87–1.00) and individual diagnoses (0.75–1.00). Concurrent validity was also established for the functional impairment/disability aspect of the MINI-KID. Sensitivity and specificity of the measure are both elevated ([Sheehan et al., 2010](#)). Others have found more moderate support for the MINI-KID, with lower sensitivity levels ([Adamowska, Adamowski, Frydecka, & Kiejna, 2014](#)).

### ***Structured Clinical Interview for DSM-IV Childhood Diagnoses***

The *Structured Clinical Interview for DSM-IV Childhood Diagnoses* (KID-SCID) was designed as a child measure derived from the SCID, a commonly used structured interview for adults ([Matzner, Silva, Silvan, Chowdhury, & Nastasi, 1997](#)). In essence, key child disorder modules were added to the original measure, notably those for attention-deficit/hyperactivity, oppositional defiant, and conduct disorders. Original modules include those for various anxiety, mood, and adjustment disorders. The interview can be administered to youth and parents. Administration time is approximately 90 min per informant.

Most of the psychometric analysis of the KID-SCID has focused on the disruptive behavior disorder categories, with initial interrater  $\kappa$  values of 0.84 for attention-deficit/hyperactivity and conduct disorders and 0.63 for oppositional defiant disorder ([Matzner et al., 1997](#)). [Smith and colleagues \(2005\)](#) performed a more extensive psychometric analysis of the measure, finding interrater  $\kappa$  values to range from 0.40 to 1.00. Values tended to be lower for substance use problems and best for attention-deficit/hyperactivity disorder. Convergent validity among the disruptive behavior disorder categories was established via correlations with reported mental health problems ([Smith, Huber, & Hall, 2005](#)). Others have found internal consistency Cronbach's  $\alpha$  values for the KID-SCID child interview to be generally

strong across most individual diagnoses, and particularly for depression (0.90), posttraumatic stress disorder (0.85), and attention-deficit/hyperactivity disorder (0.80) (Roelofs, Muris, Braet, Arntz, & Beelen, 2015). Interrater  $\kappa$  values in another study ranged from 0.79 to 1.00 (Van Vlierberghe, Braet, Goossens, & Mels, 2009).

## Specific structured interviews

As mentioned earlier, a second main type of structured and semistructured interviews for child and adolescent psychiatric disorders involves those that concentrate in more depth on one particular condition or major diagnostic category. Some of these specific structured interviews do cover many or most DSM-based psychiatric disorders, but tend to be geared towards specialized populations and research settings. An advantage of these specific interviews is that they often require less time and can be targeted more intently on youth suspected of, or referred for, a given condition.

A complete description of every specific structured interview for children and adolescents is beyond the scope of this chapter, in part because some are published with little psychometric information or are very preliminary in nature. This section thus covers prominent examples of interview measures for specific diagnostic areas, and is more illustrative than exhaustive. The reader should note as well that many rating scales for children and adolescents are purportedly useful in an interview format, but those measures are not described here.

### Anxiety disorders

The *Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions* (ADIS-C/P) (Silverman & Albano, 1996) is a venerable semistructured interview for youth aged 7–16 years, with child and parent versions. The measure was modeled after the adult version of the interview but concentrates more intently on anxiety disorders and conditions most pertinent to children and adolescents. Diagnoses are assigned based on child report, based on parent report, and based on combined report. Items are largely in yes–no format, though open-ended and other responses are allowed and recorded. Instructions are provided throughout for each diagnostic section. In addition, impairment ratings from all informants for each assigned diagnosis is derived from a 0 to 8 scale via a Feelings Thermometer. Administration time is approximately 60 min per informant, depending on the nature and severity of the presenting problems.

Like many semistructured diagnostic interviews for children and adolescents, the ADIS-C/P covers major diagnostic categories such as disruptive behavior disorders, substance use disorders, and other conditions. The large bulk of the measure, however, is devoted to detailed sections on each major anxiety disorder, especially separation anxiety, social anxiety, and generalized anxiety disorders as well as posttraumatic stress disorder, selective mutism, and specific phobia. In addition,

separate questions and sections on school refusal behavior, sleep terrors, and somatic complaints are provided for greater context. A key advantage of the measure is its utility for collecting substantial information about problems most likely referred to the clinician.

The ADIS-C/P has displayed strong  $\kappa$  values and intraclass correlation coefficients for child, parent, and combined diagnoses. For combined diagnoses across age groups, for example,  $\kappa$  values for the major anxiety disorders ranged from 0.80 to 0.92 in an initial study (Silverman, Saavedra, & Pina, 2001). Impairment ratings generally correlated with diagnoses and especially for specific phobia and generalized anxiety disorder. In addition, interrater agreement on composite diagnoses has been found to range from 0.82 to 1.00 for the anxiety disorders in the measure, and are slightly better for children than adolescents.  $\kappa$  Values for the disruptive behavior disorders tend to be lower, though this could be due to the expected lower prevalence of these disorders among youth presenting for anxiety conditions (Lyneham, Abbott, & Rapee, 2007). Others have found strong support for the concurrent validity of the ADIS-C/P with various rating scales (Wood, Piacentini, Bergman, McCracken, & Barrios, 2002). The ADIS-C/P remains the gold standard for assessing anxiety-related disorders in children and adolescents and is frequently used in clinical research settings.

### **Trauma-related conditions**

Other interviews have been specifically developed for trauma-related conditions in children and adolescents. This is particularly important given rapid changes in understanding the developmentally different nature of trauma symptoms across childhood and adolescence, a fact now represented in most taxonomic nomenclatures. The *Child PTSD Symptom Scale for DSM-5 for Trauma-Exposed Children and Adolescents* (CPSS-5), for example, is fitted to the new developmental criteria for posttraumatic stress disorder (Foa, Asnaani, Zang, Capaldi, & Yeh, 2018). The interview version of this measure (CPSS-5-I) contains 27 items surrounding DSM-5 diagnostic criteria as well as symptom severity in the past month. Subscale scores for intrusion, avoidance, cognition and mood changes, and increased arousal and reactivity can be derived. Internal consistency Cronbach's  $\alpha$  values for the subscales range from 0.67 to 0.81 and test-retest reliability ranges from 0.76 to 0.86. Convergent and discriminant validity have been established as well.

Other semistructured interviews relevant to this area include the *Developmental Trauma Disorder Semistructured Interview* and *Traumatic Experiences Screening Instrument* (Ford, Spinazzola, van der Kolk, & Grasso, 2018). The former examines developmental trauma symptom sets related to emotion/somatic, attentional/behavioral, and interpersonal/self-dysregulation areas of functioning. The latter examines specific types of trauma and interpersonal victimization. In addition, others have derived an acute stress module interview for youth from the DICA (Kassam-Adams et al., 2013). The *UCLA Child/Adolescent PTSD Reaction Index for DSM-5* is a semistructured interview that covers a child's trauma history as well as DSM-5

criteria for PTSD among school-age children and adolescents. Child and parent versions of this interview are available ([Pynoos & Steinberg, 2014](#)).

### **Selective mutism**

Two general interviews have been designed for selective mutism, a condition in which children decline or fail to speak in public situations despite speaking well at home. The *Functional Diagnostic Protocol* covers conditions under which selective mutism occurs and what reinforcers maintain mutism over time ([Schill, Kratochwill, & Gardner, 1996](#)). The interview also helps clinicians gather information about factors that may impact a child's selective mutism, including psychosocial and physical events, affective states, personality traits, cognitive variables, skills deficits, and setting events.

In addition, the *Selective Mutism Comprehensive Diagnostic Questionnaire* focuses on settings in which selective mutism occurs as well as frequency, severity, and pervasiveness of symptoms ([Mulligan, Hale, & Shipon-Blum, 2015](#)). Items surround socialization variables, peer interactions, communication methods, developmental history, school functioning, coexisting disorders, and a child's personality, body language, and behavior. Parents can also rate intensity of various mutism behaviors on a 1–10 scale. The measure reportedly has strong content validity, but additional psychometric studies of both selective mutism interviews remain needed.

### **Other areas**

Other forms of child psychopathology, or aspects related to diagnostic conditions in children, have been assessed via specialized structured interviews as well. With respect to attention-deficit/hyperactivity disorder, for example, [Holmes and colleagues \(2004\)](#) developed a structured teacher telephone interview that covers symptoms of the disorder over the past 3 months. The *Child Attention-Deficit Hyperactivity Disorder Teacher Telephone Interview* (CHATTI) takes approximately 15–20 min to complete and includes 18 ICD and DSM items regarding inattention, hyperactivity, and impulsivity. The interview was shown to have strong interrater and test–retest reliability. Others have also formulated clinical interviews for specific intervention protocols related to attention-deficit/hyperactivity disorder ([Barkley, 1997](#)).

Child depression is often covered in the major diagnostic interviews discussed earlier in the chapter, especially because depression is one of the more reliably diagnosed disorders among these measures. However, commonly used screening devices are sometimes used in interview format and have good psychometric strength ([Stockings et al., 2015](#)). In addition, the *Children's Attributional Style Interview* (CASI) is an interactive measure of 16 positive and negative events presented to children who then provide attributions based on the dimensions of internality, stability, and globality. The interview is developmentally oriented and has demonstrated moderately acceptable reliability and validity ([Conley, Haines, Hilt, & Metalsky, 2001](#)).

Other specialized interviews continue to be developed for more circumscribed populations, a trend expected to continue in this area. For example, [Koo, Han, Park, and Kwon \(2017\)](#) developed a structured clinical interview for gambling problems in adolescents. Others have also developed structured interviews for specific anxiety disorders ([Popp, Neuschwander, Mannstadt, In-Albon, & Schneider, 2017](#)), autism ([Bishop et al., 2017](#)), preschoolers ([Olino, Dougherty, Bufferd, Carlson, & Klein, 2014](#)), and reactive attachment and disinhibited social engagement disorder ([Lehmann et al., 2018](#)).

## **Strengths and limitations of structured interviews**

Structured and semistructured interviews have several advantages with respect to assessment of child and adolescent psychiatric disorders. Advantages include standardization of administration and scoring, applicability to large clinical research settings, utility for epidemiological and treatment outcome studies, and opportunities for some flexibility among many interviews. Diagnostic clarity, especially in cases of complex comorbidity, may be more achievable with structured/semistructured than unstructured interviews as well ([Leffler, Riebel, & Hughes, 2015](#)).

Structured and semistructured interviews also have several disadvantages with respect to assessment of child and adolescent psychiatric disorders. Like all measurement endeavors, clinical child interviews are built on certain assumptions. Descriptive interviews of children are only valid if the underlying theory for the interview is correct. Clinical diagnostic interviews assume both that the taxonomy of diagnoses is correct (most often the DSM system) and that the most discriminating symptoms are included in the interview. However, clinical child interviews also rely more explicitly on additional assumptions. First, informants are assumed to be able to provide valid information about a child's emotional, behavioral, and social functioning. For example, parents and teachers tend to identify externalizing problems (e.g., attention, alcohol and drug use, stealing) at higher rates than child self-report. In contrast, child self-report is often assumed to better for internalizing problems (e.g., fears, worries, and anxieties). Similarly, matching information to informants is critical (e.g., should a school teacher have direct knowledge of a child's sleep habits). Second, child clinical interviews must be developmentally appropriate. For example, children tend to struggle more with time and frequency of past events as well as duration and onset of current problems ([Fallon & Schwab-Stone, 1994](#)). Of note, the reliability of clinical child interviews increases rapidly with age such that adolescents show reliability typical of adults ([Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985](#)). Thus, interviews can be subject to bias, rely heavily on DSM categories and less on dimensional approaches to psychopathology, and may neglect substantial and pertinent historical and contextual information ([Frick, Barry, & Kamphaus, 2010](#)).

It is also the case that most take a long time to revalidate once diagnostic criteria change and they are generally quite lengthy, making them less amenable to most

clinical settings where diagnoses are used for billing purposes, there are constraints on resources and efficiency is highly valued. In addition,  $\kappa$  values for most diagnostic categories derived from structured interviews remain in the 0.40–0.60 range, meaning that shorter measures or rating scales could provide comparable diagnostic information for much less time and cost (Boyle et al., 2017). Despite these assumptions and limitations, child clinical interviewing underlies a substantial portion of our scientific evidence regarding disorders of childhood and adolescence.

## Conclusion

Structured and semistructured diagnostic interviews for children and adolescents have been an important and crucial part of assessment in clinical child psychology and psychiatry for decades. The measures have contributed immensely to our understanding of diagnostic prevalence, comorbidity patterns, subtypes, and informant variance. As the field progresses towards more nuanced understandings of many mental disorders in youth, however, interviews are expected to become more nimble, brief, precise, and targeted towards a specific problem. In addition, greater research is needed to explore the incremental utility of interviews compared to screening devices, rating scales, questionnaires, observations, and general behavioral assessment.

## References

- Adamowska, S., Adamowski, T., Frydecka, D., & Kiejna, A. (2014). Diagnostic validity polish language version of the questionnaire MINI-KID (Mini International Neuropsychiatry Interview for Children and Adolescent). *Comprehensive Psychiatry*, 55, 1744–1750.
- Ambrosini, P. J. (2000). Historical development and present status of the Schedule for Affective Disorders and Schizophrenia for school-age children (K-SADS). *Journal of the American Academy of Child and Adolescent Psychiatry*, 39, 49–58.
- Angold, A., & Costello, E. J. (2000). The child and adolescent psychiatric assessment (CAPA). *Journal of the American Academy of Child and Adolescent Psychiatry*, 39, 39–48.
- Angold, A., Prendergast, M., Cox, A., Harrington, R., Simonoff, E., & Rutter, M. (1995). The child and adolescent psychiatric assessment (CAPA). *Psychological Medicine*, 25, 739–753.
- Barkley, R. A. (1997). *Defiant children: A clinician's manual for assessment and parent training* (2nd ed). New York, NY: Guilford.
- Bishop, S. L., Huerta, M., Gotham, K., Alexandra Haydahl, K., Pickles, A., Duncan, A., ... Lord, C. (2017). The autism symptom interview, school-age: A brief telephone interview to identify autism spectrum disorders in 5- to 12-year-old children. *Autism Research*, 10, 78–88.
- Boyle, M. H., Duncan, L., Georgiades, K., Bennett, K., Gonzalez, A., Van Lieshout, R. J., ... Lipman, E. L. (2017). Classifying child and adolescent psychiatric disorder by problem checklists and standardized interviews. *International Journal of Methods in Psychiatric Research*, 26, e1544.

- Chambers, W. J., Puig-Antich, J., Hirsch, M., Paez, P., Ambrosini, P. J., Tabrizi, M. A., ... Davies, M. (1985). The assessment of affective disorders in children and adolescents by semistructured interview: Test-retest reliability of the Schedule for Affective Disorders and Schizophrenia for School-Age Children, Present Episode Version. *Archives of General Psychiatry*, 42(7), 696–702. Available from <https://doi.org/10.1001/archpsyc.1985.01790300064008>.
- Conley, C. S., Haines, B. A., Hilt, L. M., & Metalsky, G. I. (2001). The Children's Attributional Style Interview: Developmental tests of cognitive diathesis-stress theories of depression. *Journal of Abnormal Child Psychology*, 29, 445–463.
- 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(1), 265–275. Available from <https://doi.org/10.2307/1130193>.
- Egger, H. L., Erkanli, A., Keeler, G., Potts, E., Walter, B. K., & Angold, A. (2006). Test-retest reliability of the preschool age psychiatric assessment (PAPA). *Journal of the American Academy of Child & Adolescent Psychiatry*, 45(5), 538–549. Available from <https://doi.org/10.1097/01.chi.0000205705.71194.b8>.
- Elkind, D. (1964). Piaget's semi-clinical interview and the study of spontaneous religion. *Journal for the Scientific Study of Religion*, 4(1), 40–47. Available from <https://doi.org/10.2307/1385202>.
- Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The schedule for affective disorders and schizophrenia. *Archives of General Psychiatry*, 35(7), 837–844. Available from <https://doi.org/10.1001/archpsyc.1978.01770310043002>.
- Fallon, T., & Schwab-Stone, M. (1994). Determinants of reliability in psychiatric surveys of children aged 6–12. *Child Psychology & Psychiatry & Allied Disciplines*, 35(8), 1391–1408. Available from <https://doi.org/10.1111/j.1469-7610.1994.tb01282.x>.
- Fisher, P., Lucas, L., Lucas, C., Sarsfield, S., & Shaffer, D. (2006). *Interviewer manual*. New York: Columbia University DISC Development Group.
- Foa, E. B., Asnaani, A., Zang, Y., Capaldi, S., & Yeh, R. (2018). Psychometrics of the Child PTSD Symptom Scale for DSM-5 for trauma-exposed children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, 47, 38–46.
- Ford, J. D., Spinazzola, J., van der Kolk, B., & Grasso, D. J. (2018). Toward an empirically based developmental trauma disorder diagnosis for children: Factor structure, item characteristics, reliability, and validity of the Developmental Trauma Disorder Semi-Structured Interview. *Journal of Clinical Psychiatry*, 79, 17m11675.
- Frick, P. J., Barry, C. T., & Kamphaus, R. W. (2010). *Clinical assessment of child and adolescent personality and behavior* (3rd ed). New York: Springer.
- Graham, P., & Rutter, M. (1968). The reliability and validity of the Psychiatric Assessment of the Child: II Interview with the parent. *The British Journal of Psychiatry*, 114(510), 581–592. Available from <https://doi.org/10.1192/bjp.114.510.581>.
- Herjanic, B., & Reich, W. (1982). Development of a structured psychiatric interview for children: Agreement between child and parent on individual symptoms. *Journal of Abnormal Child Psychology*, 10(3), 307–324. Available from <https://doi.org/10.1007/BF00912324>.
- Holmes, J., Lawson, D., Langley, K., Fitzpatrick, H., Trumper, A., Pay, H., ... Thapar, A. (2004). The child attention-deficit hyperactivity disorder teacher telephone interview (CHATTI): Reliability and validity. *British Journal of Psychiatry*, 184, 74–78.
- Jones, K. D. (2010). The unstructured clinical interview. *Journal of Counseling & Development*, 88(2), 220–226. Available from <https://doi.org/10.1002/j.1556-6678.2010.tb00013.x>.

- Kassam-Adams, N., Gold, J. I., Montaño, Z., Kohser, K. L., Cuadra, A., Muñoz, C., ... Armstrong, F. D. (2013). Development and psychometric evaluation of child acute stress measures in Spanish and English. *Journal of Traumatic Stress, 26*, 19–27.
- Kaufman, J., & Schweder, A. E. (2004). The Schedule for Affective Disorders and Schizophrenia for School-Age Children: Present and Lifetime version (K-SADS-PL). In M. J. Hilsenroth, & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment* (pp. 247–255). Hoboken, NJ: John Wiley & Sons.
- Koo, H. J., Han, D. H., Park, S. Y., & Kwon, J. H. (2017). The structured clinical interview for DSM-5 Internet gaming disorder: Development and validation for diagnosing IGD in adolescents. *Psychiatry Investigation, 14*, 21–29.
- Kraepelin, E. (1921). In R. M. Barclay (Ed.), *Manic-depressive insanity and paranoia*. Edinburgh: Livingstone, Trans.
- Lauth, B., Arnelsson, G. B., Magnússon, P., Skarphéðinsson, G. Á., Ferrari, P., & Pétrusson, H. (2010). Validity of K-SADS-PL (Schedule for Affective Disorders and Schizophrenia for School-Age Children—Present and Lifetime version) depression diagnoses in an adolescent clinical population. *Nordic Journal of Psychiatry, 64*, 409–420.
- Leffler, J. M., Riebel, J., & Hughes, H. M. (2015). A review of child and adolescent diagnostic interviews for clinical practitioners. *Assessment, 22*, 690–703.
- Lapouse, R., & Monk, M. A. (1959). Fears and worries in a representative sample of children. *American Journal of Orthopsychiatry, 29*(4), 803–818. Available from <https://doi.org/10.1111/j.1939-0025.1959.tb00250.x>.
- Lehmann, S., Monette, S., Egger, H., Breivik, K., Young, D., Davidson, C., & Minnis, H. (2018). Development and examination of the Reactive Attachment Disorder and Disinhibited Social Engagement Disorder Assessment Interview. *Assessment, 2018*, 1–17.
- Lyneham, H. J., Abbott, M. J., & Rapee, R. M. (2007). Interrater reliability of the Anxiety Disorders Interview Schedule for DSM-IV: Child and parent version. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*, 731–736.
- Matzner, F., Silva, R., Silvan, M., Chowdhury, M., & Nastasi, L. (1997). Preliminary test-retest reliability of the KID-SCID. *Scientific Proceedings, American Psychiatric Association Meeting*, Washington, DC.
- McConaughy, S. H. (2013). *Clinical interviews for children and adolescents: Assessment to intervention* (2nd ed). New York, NY: Guilford Press.
- Merrell, K. W. (2008). *Behavioral, social, and emotional assessment of children and adolescents* (3rd ed). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Mulligan, C. A., Hale, J. B., & Shipon-Blum, E. (2015). Selective mutism: Identification of subtypes and implications for treatment. *Journal of Education and Human Development, 4*, 79–96.
- Olino, T. M., Dougherty, L. R., Bufferd, S. J., Carlson, G. A., & Klein, D. N. (2014). Testing models of psychopathology in preschool-aged children using a structured interview-based assessment. *Journal of Abnormal Child Psychology, 42*, 1201–1211.
- Popp, L., Neuschwander, M., Mannstadt, S., In-Albon, T., & Schneider, S. (2017). Parent-child diagnostic agreement on anxiety symptoms with a structured diagnostic interview for mental disorders in children. *Frontiers in Psychology, 8*, 404.
- Pynoos, R. S., & Steinberg, A. M. (2014). *UCLA child/adolescent PTSD reaction index for DSM-5*. Los Angeles, CA: University of California at Los Angeles.
- Reich, W. (2000). Diagnostic interview for children and adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 59–66.

- Robins, L. N., Helzer, J. E., Croughan, J. L., & Ratcliff, K. S. (1981). National Institute of Mental Health diagnostic interview schedule: Its history, characteristics, and validity. *Archives of General Psychiatry*, 38(4), 381–389. Available from <https://doi.org/10.1001/archpsyc.1981.01780290015001>.
- Roelofs, J., Muris, P., Braet, C., Arntz, A., & Beelen, I. (2015). The Structured Clinical Interview for DSM-IV childhood diagnoses (Kid-SCID): First psychometric evaluation in a Dutch sample of clinically referred youths. *Child Psychiatry and Human Development*, 46, 367–375.
- Rosen, B. M., Bahn, A. K., & Kramer, M. (1964). Demographic and diagnostic characteristics of psychiatric clinic outpatients in the USA, 1961. *American Journal of Orthopsychiatry*, 34(3), 455–468. Available from <https://doi.org/10.1111/j.1939-0025.1964.tb02214.x>.
- Rourke, K. M., & Reich, W. (2004). The diagnostic interview for children and adolescents (DICA). In M. J. Hilsenroth, & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment: Vol. 2. Personality assessment* (pp. 271–280). Hoboken, NJ: John Wiley & Sons.
- Rutter, M., & Graham, P. (1968). The reliability and validity of the psychiatric assessment of the child: I Interview with the child. *The British Journal of Psychiatry*, 114(510), 563–579. Available from <https://doi.org/10.1192/bjp.114.510.563>.
- Sattler, J. M. (2008). *Assessment of children: Cognitive foundations, 5th edition* (5th ed.). San Diego: Jerome M. Sattler, Publisher.
- Schill, M. T., Kratochwill, T. R., & Gardner, W. I. (1996). An assessment protocol for selective mutism: Analogue assessment using parents as facilitators. *Journal of School Psychology*, 34, 1–21.
- Shaffer, D., Fisher, P., & Lucas, C. (2004). The diagnostic interview schedule for children (DISC). In M. J. Hilsenroth, & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment: Vol. 2. Personality assessment* (pp. 256–270). Hoboken, NJ: John Wiley & Sons.
- 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.
- Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, 41(5), 551–561. Available from <https://doi.org/10.1002/pits.10176>.
- Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., ... Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, 59(Suppl. 20), 22–33.
- Sheehan, D. V., Sheehan, K. H., Shytle, R. D., Janavs, J., Bannon, Y., Rogers, J. E., ... Wilkinson, B. (2010). Reliability and validity of the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI-KID). *Journal of Clinical Psychiatry*, 71, 313–326.
- Silverman, W. K., & Albano, A. M. (1996). *The Anxiety Disorders Interview Schedule for Children for DSM-IV, child and parent versions*. San Antonio, TX: Psychological Corporation.
- Silverman, W. K., & Nelles, W. B. (1988). The Anxiety Disorders Interview Schedule for Children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 27(6), 772–778. Available from <https://doi.org/10.1097/00004583-198811000-00019>.

- Silverman, W. K., Saavedra, L. M., & Pina, A. A. (2001). Test-retest reliability of anxiety symptoms and diagnoses with the Anxiety Disorders Interview Schedule for DSM-IV: Child and parent versions. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 937–944.
- Smith, D. C., Huber, D. L., & Hall, J. A. (2005). Psychometric evaluation of the Structured Clinical Interview for DSM-IV childhood diagnoses (KID-SCID). *Journal of Human Behavior in the Social Environment, 11*, 1–21.
- Stockings, E., Degenhardt, L., Lee, Y. Y., Mihalopoulos, C., Liu, A., Hobbs, M., & Patton, G. (2015). Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders, 174*, 447–463.
- Van Vlierberghe, L., Braet, C., Goossens, L., & Mels, S. (2009). Psychiatric disorders and symptom severity in referred versus non-referred overweight children and adolescents. *European Child and Adolescent Psychiatry, 18*, 164–173.
- Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*(1), 54–60. Available from <https://doi.org/10.1037/0735-7028.26.1.54>.
- Weller, E. B., Weller, R. A., Fristad, M. A., & Rooney, M. T. (1999). *ChIPS: Children's interview for psychiatric syndromes*. Washington, DC: American Psychiatric Publishing.
- Weller, E. B., Weller, R. A., Fristad, M. A., Rooney, M. T., & Schechter, J. (2000). Children's interview for psychiatric syndromes (ChIPS). *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 76–84.
- Wood, J. J., Piacentini, J. C., Bergman, R. L., McCracken, J., & Barrios, V. (2002). Concurrent validity of the anxiety disorders section of the Anxiety Disorders Interview Schedule for DSM-IV: Child and parent versions. *Journal of Clinical Child and Adolescent Psychology, 31*, 335–342.

# Diagnostic and symptom interviews for adults

12

*Daniel N. Allen and Megan L. Becker*

Department of Psychology, University of Nevada, Las Vegas, NV, United States

## Introduction

The clinical interview has numerous goals as discussed in the Clinical Interviewing chapter of this handbook, but two objectives are to: (1) determine the presence and severity of psychiatric symptoms, and (2) establish a diagnosis that will be the focus of treatment. Interviews were developed for both purposes, with a number of structured and semistructured interviews currently available to establish psychiatric diagnosis, along with many symptom or behavior ratings scales to establish the presence and severity of psychiatric symptoms. These interview procedures were developed to improve reliability of observations made by clinicians and researchers. Interviews like the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnoses (SCID) or the Diagnostic Interview Schedule (DIS) help increase reliability in diagnoses, and a major impetus for their development was the acceptance of the DSM diagnostic system, most notably the DSM-III and later versions. Broad acceptance of a unified diagnostic system provided clear criteria which could be systematically addressed using structured and semistructured interview formats.

In addition to structured and semistructured interviews like the SCID or DIS, the primary purpose for both of which is to establish a psychiatric diagnosis, common symptom and behavior rating scales quantify current severity levels for a variety of psychiatric symptoms. These scales have both clinical and research applications. Symptom rating scales are essential for determining changes in symptoms in response to pharmacological challenge, and often provide the basis for judging effectiveness of new treatments. Ratings may also be used descriptively, to establish the presence of a specific subtype or syndrome (e.g., deficit syndrome schizophrenia), or in correlational or covariance analyses. In clinical settings, ratings scales are used to document current symptom severity (e.g., past week) or over longer time frames (1 month), depending on the purpose of the evaluation. Ratings help establish baseline levels of psychiatric symptoms at initial intake and can then be repeatedly administered to judge responses to treatment over weeks or months.

They may also be used as an endpoint to the treatment process to demonstrate treatment effectiveness. Although the ratings may be useful for informing the diagnostic process, the rating scales themselves are not designed to provide a formal diagnosis.

The following sections contain a review of some of the more common current diagnostic interviews and symptoms rating scales. The coverage is not exhaustive, but is meant to familiarize the reader with scales that are most widely used and have particular utility in clinical and research settings.

## Diagnostic Interviews

As discussed in the Clinical Interviewing chapter of this handbook, development of fully structured and semistructured diagnostic interviews began in earnest following the publication of the DSM-III ([APA, 1980](#)), because the diagnostic criteria made necessary the gathering uniform and detailed data to support differential diagnoses. At the forefront of this development was Spitzer and colleagues' Research Diagnostic Criteria and the Schedule for Affective Disorders ([Endicott & Spitzer, 1978](#); [Spitzer, Endicott, & Robins, 1978](#)). Their work and the work of others focused on standardizing and scheduling the diagnostic interview to address matters related to unreliable diagnosis resulting primarily from criterion and information variance which limited the usefulness of psychiatric diagnoses for prevention efforts, prescribing treatment, and stimulating research. The procedures they developed and those that followed, reduced diagnostic error resulting from criterion and information variance, which contributed to the continued development of structured and semistructured interviews and their increased use, particularly in research settings. An ongoing need persists to develop structured and semistructured interviews which are simpler and easier to navigate, thereby affording greater practical application in clinical settings where diagnostic reliability continues to pose challenges.

The following sections include a review of the most popular structured and semi-structured diagnostic interviews, focusing on those that are more comprehensive in scope and assess a broad range of DSM disorders. There are also more specialized interview procedures available to assist in making specific diagnoses, such as the Autism Diagnostic Interview—Revised ([Rutter, Le Couteur, & Lord, 2003](#)), or the Eating Disorders Examination ([Fairburn & Cooper, 1993](#)). Given the scope of the chapter, we will not review these more specialized interview procedures, but many are available with strong psychometric support and should be considered when a comprehensive diagnostic evaluation is not required. The following structured and semistructured interviews are reviewed:

- Structured Clinical Interview for the DSM-5 (SCID-5)
- Mini International Neuropsychiatric Interview (MINI)
- The Diagnostic Interview Schedule (DIS)
- Composite International Diagnostic Interview (CIDI)
- International Personality Disorder Examination (IPDE)
- Structured Clinical Interview for the DSM-5 Personality Disorders (SCID-5-PD)

## **Structured Clinical Interview for the DSM-5**

The Structured Clinical Interview for DSM-5 Disorders (SCID-5) is widely considered the “gold standard” for psychiatric diagnoses. In its many forms, the SCID is perhaps the most widely used diagnostic research instrument in the English language. It offers a thorough assessment of diagnosis for a multitude of disorders and is modularized so it can be tailored to fit the scope of the research question. This semistructured interview is designed for use with adults who have an eighth grade or higher reading level. The SCID may be used to assess psychiatric and general medical patients, and individuals from the community for whom no diagnosis is identified. There are several versions of the SCID developed for use in research and clinical settings. The SCID-5 Research Version (SCID-5-RV; [First, Williams, Karg, & Spitzer, 2015a](#)) is broadest in scope, and allows subtype specification, severity, and course specifiers. The SCID-5-RV is a long, in-depth interview designed to screen for most DSM-5 ([APA, 2013](#)) clinical diagnoses (AXIS I disorders in the DSM-IV). There is a standard “core” configuration that includes disorders routinely assessed in most research studies, and an “enhanced” configuration that includes all core configuration disorders and numerous optional disorders (e.g., separation anxiety disorder, insomnia disorder, body dysmorphic disorder, somatic symptom disorder, gambling disorder). The SCID-5 Clinical Version (SCID-5-CV; [First, Williams, Karg, & Spitzer, 2015b](#)) is a shortened version, which allows for a structured diagnostic interview format for use in clinical settings. Questions and scoring are included in a single administration booklet with an accompanying single-use score sheet for recording responses.

The full SCID versions may be inefficient as a broad-range diagnostic tool outside of a research environment. A thorough SCID interview by an experienced clinician will take longer than an hour for most clinical cases. The DSM-5 version provides a comprehensive screening module, which may serve to ultimately reduce total administration time. Also, subsections of the SCID may be administered in a “stand-alone” fashion when specific diagnostic concerns necessitate diagnosis of a particular disorder. Despite this, it remains unlikely that the SCID will become practical for most assessments outside of the research setting. A version of the SCID for clinical trials (SCID-5-CT; [First, Williams, Karg, & Spitzer, 2015c](#)) also exists which can be tailored to screen for study inclusion and exclusion criteria. SCID user’s guides and other training materials are available.

Establishing the validity of the SCID is challenging, because as the “gold standard” for diagnosis of many DSM disorders, it typically requires comparison to other diagnostic interviews which are less comprehensive and may be less reliable or valid (see discussion of the MINI for DSM-5) ([Sheehan et al., 1998, 2015](#)). Evidence indicates that the SCID-5 showed improved validity for dimensional severity of diagnoses as classified in the DSM-5 beyond that of categorical diagnoses ([Shankman et al., 2018](#)). Additionally, it demonstrated good predictive and concurrent validity ([Shankman et al., 2018](#)).

Predating the SCID-5, the SCID-IV had acceptable to excellent interrater reliability, with adequate administrator training ([Lobbestael, Leurgans, & Arntz, 2011](#)).

Although information on the interrater reliability of the SCID-5 is limited, agreement between raters on the somatic symptom disorder and illness anxiety disorder diagnoses was reportedly moderate to nearly perfect (overall  $\kappa = .85$ ) (Axelsson, Andersson, Ljótsson, Wallhed Finn, & Hedman, 2016). Indeed, many sites require a SCID rater to achieve consecutive 100% concurrence with an experienced interviewer for primary diagnosis. Additionally, Regier et al. (2013) examined test-retest reliability and concluded that it was very good ( $\kappa = 0.60\text{--}0.79$ ) for posttraumatic stress disorder (PTSD), complex somatic symptom disorder, and major neurocognitive disorder, good ( $\kappa = 0.40\text{--}0.59$ ) for schizophrenia, schizoaffective disorder, bipolar I disorder, binge eating disorder, alcohol use disorder, mild neurocognitive disorder, and borderline personality disorder, and questionable ( $\kappa = 0.20\text{--}0.39$ ) for major depressive disorder, generalized anxiety disorder, mild traumatic brain injury, and antisocial personality disorder.

### ***Mini International Neuropsychiatric Interview***

As a diagnostic screening tool, the clinician-rated MINI is among the most widely used psychiatric structured diagnostic interviews. It is a quick, robust tool for the assessment of current, past, and lifetime diagnoses of common International Classification of Diseases (ICD-10) and DSM-5 disorders. The MINI 7.0.2. is the latest version, as it corresponds with diagnostic criteria from the DSM-5. A series of validation studies with a prior version of the MINI (Sheehan et al., 1998, 2015) revealed high negative predictive value ( $>0.92$ ), for all assessed diagnostic categories when compared with the SCID, and similarly  $>0.88$  for the CIDI for the ICD-10 (Kessler & Üstün, 2004). This instrument carries some advantages, such as that it requires substantially less time to administer than the SCID and the CIDI, its prior versions demonstrated good to excellent reliability for most diagnostic criteria, it was translated into more than 30 languages, and it requires less training to administer than many structured and semistructured interviews such as the SCID. When used in the Iowa State prison setting, two 120-minute training sessions were sufficient for those with an undergraduate level education (Black, Arndt, Hale, & Rogerson, 2004; Gunter et al., 2008).

Sheehan et al. (1997) assessed the validity of the MINI in relation to other gold standard structured diagnostic systems, the SCID and CIDI. This well-cited study indeed makes a strong case for the diagnostic reliability of this measure in general, although the exceptionally homogeneous sample (96% Caucasian) has unknown implications for validity of the MINI in other racial and ethnic groups. Upon validation, the MINI demonstrated good interrater and test-retest reliability for the original sample of participants (Sheehan et al., 1997, 1998). Further, good to excellent reliability coefficients were obtained in all studies noted above.

### ***The Diagnostic Interview Schedule***

The Diagnostic Interview Schedule version IV (DIS-IV) (Robins, Cottler, Bucholz, & Compton, 1995) evaluates current and lifetime presence of DSM-IV mental

disorders. The DIS has not been updated for DSM-5 criteria which limits its application in clinical practice and research settings to the extent that the diagnostic criteria for the disorder of interest changed substantially from the DSM-IV to DSM-5. The DIS is a fully structured diagnostic interview that is organized in 19 diagnostic modules that cover most Axis I disorders.

Diagnoses include substance abuse and dependence, schizophrenia, mood disorders, anxiety disorders, and a small selection of other disorders, such as those originating in childhood. Professionals and nonprofessionals can administer the DIS provided that they complete the required training, which is approximately 1 week long. Administration time is approximately 2 hours, during which questions are read verbatim and no opportunity is given for unstructured questions. A probe flow chart shows how optional probes should be used to obtain accurate clinical ratings. A training manual is included that describes how to reliably code the clinical ratings of specific items (Robins, Cottler, & Keating, 1991). Items are scored in a format that combines clinical relevance and possible etiology (1 = denial of symptom; 2 = subclinical; 3 = clinically relevant symptoms, with etiology caused by medications, drugs, or alcohol; 4 = clinically relevant symptoms, with etiology caused by physical illness or injury; 5 = clinically relevant symptoms, with etiology caused by psychiatric disorder). Additionally, interviewers' ratings are based on the onset, duration, and recency of symptoms. The DIS is structured to elucidate information regarding organic etiology (exogenous substances, medical conditions) and includes an assessment of cognitive functioning via the Mini Mental Status Examination (MMSE; Folstein, Folstein, & McHugh, 1975). The DIS also allows researchers to conduct comparisons across diagnostic systems because it retained criteria from other systems, such as older DSM versions, Feighner criteria (Feighner et al., 1972), and Research Diagnostic Criteria (Spitzer et al., 1978) (Rogers, 2001).

The DIS was designed as a research tool for large epidemiological studies (Epidemiologic Catchment Area Program) to assess the prevalence and incidence of mental disorders in the United States (Regier et al., 1984). Several computerized versions of the DIS exist, which include patient and interviewer-based administrations. However, these versions typically do not cover as many diagnoses as the DIS-IV. In addition, there is a shortened paper and pencil version that can be self-administered and covers depressive, anxiety, and alcohol disorders (Kovess & Fournier, 1990). The DIS-IV was translated to several languages, including Spanish and Chinese. The DIS also has well-established validity and reliability. Notably, many of the reliability and validity studies in the literature were conducted using earlier versions of the DIS and their corresponding diagnostic system, so such results should be interpreted with this knowledge. Overall, research shows the diagnostic reliability and validity of the DIS is moderate to good (e.g., Helzer et al., 1985; Hesselbrock, Stabenau, Hesselbrock, Mirkin, & Meyer, 1982; North et al., 1997; Wells, Burnam, Leake, & Robins, 1988). Limitations of the DIS include an increased focus on etiology rather than symptom severity, emphasis on diagnosis over symptom evaluation, and additionally, research suggests that it is vulnerable to response styles (Alterman et al., 1996; Cottler, Compton, Ridenour, Abdahlla, & Gallagher, 1998; Rogers, 2001). However, some of the main advantages of the DIS

are its utility in screening large samples of people for undetected mental disorders (epidemiological research), its allowance for administration by nonprofessionals, and the fact that it was translated and validated in several languages. In addition, there is a children's version of the DIS, the Diagnostic Interview Schedule for Children (DISC; [Columbia DISC Development Group, 1999](#); [National Institute of Mental Health, 1991](#)).

### ***Composite International Diagnostic Interview***

[Kessler and Üstün \(2004\)](#) developed the CIDI under the sponsorship of the WHO. The CIDI has not been updated to reflect DSM-5 criteria, thus its validity is likely questionable for diagnoses that had changes in criteria between the DSM-IV and DSM-5. Based on the DIS and expanded with questions from the Present State Examination (PSE), the CIDI has several modified items to extend its utility internationally. The CIDI's purpose was to facilitate cross-cultural epidemiologic and comparative studies, so it is a highly structured interview. It can be administered by nonprofessionals after extensive training and is easily translated to different languages. In addition, the CIDI provides both DSM and ICD diagnoses. Administration time can vary from 1 hour and 15 minutes to 1 hour and 45 minutes, depending on the experience of the interviewer. The most current version of the CIDI (3.0) is composed of 42 sections that assess a wide variety of disorders included in the DSM-IV and ICD-10. The main sections include mood, anxiety, impulse control, and substance abuse disorders among others. There is also a computer version available, the CIDI 3.0 Computer Assisted Personal Interview (CAPI V21.1.3). The CIDI is translated to several languages. Other versions of the CIDI that are available are the CIDI Primary Health Care Version that was developed to address psychological problems frequently observed in medical settings ([Janca et al., 1994](#)) and the University of Michigan version of the CIDI (UM-CIDI) that excludes the MMSE and somatoform disorders and includes auditory processing disorder and PTSD ([Rogers, 2001](#)). Generally, the reliability and validity of the CIDI is moderate to high ([Cottler et al., 1997](#); [Haro et al., 2006](#); [Janca et al., 1992](#); [Cooper, Peters, & Andrews, 1998](#); [Wittchen, 1994](#)). However, these studies should be interpreted carefully because several are inconsistent in terms of CIDI version, language, and diagnostic framework used. Nevertheless, this is expected considering the cross-cultural focus of the CIDI and the multiple versions that are available.

### ***International Personality Disorder Examination***

The IPDE ([Loranger et al., 1994](#)) is a semistructured interview for DSM-IV axis II disorders and is normed for use with the ICD-10. However, the IPDE has not been updated to include DSM-5 criteria and its validity likely does not extend to diagnosis with updated criteria in the DSM-5. It was created through a joint project between the WHO and the National Institutes of Health to provide well-validated research instruments for worldwide use. The instrument was created in

English and translated into a wide variety of other languages. This instrument also has the advantage of being accepted within the research community and was normed on a large sample of over 700 people. Although this measure can be time consuming, it may be administered in more than one session, provided a given diagnostic category is completed prior to discontinuation.

The IPDE was normed for adults without severe psychopathology, and those of “below-normal intelligence” (Loranger, Janca, & Sartorius, 1997). Clearly these parameters limit the population for which this interview is considered acceptable. As with other interviews, the administration time for the IPDE may be long, and in complicated cases could take over an hour and a half to administer.

### ***Structured Clinical Interview for the DSM-5 Personality Disorders***

For the DSM-IV (APA, 2013) and earlier DSM’s which used the multiaxial diagnostic system, the SCID was divided into two diagnostic interviews, one for diagnosing Axis I disorders and another for diagnosing Axis II personality disorders, the latter referred to as the SCID-II. The DSM-5 excluded the multiaxial diagnostic system, which was a substantial revision, and included an alternative model for diagnosing personality disorders. The SCID-II was revised to the SCID-5 Personality Disorders (SCID-5-PD; First, Williams, Benjamin, & Spitzer, 2016) and the SCID-5 Alternative Model for Personality Disorders (SCID-5-AMPD; First, Skodol, Bender, & Oldham, 2018) to address these significant changes.

The SCID-5-PD allows for categorical or dimensional diagnosis of the 10 DSM-5 personality disorders (Paranoid, Schizoid, Schizotypal, Antisocial, Borderline, Histrionic, Narcissistic, Avoidant, Dependent, Obsessive—Compulsive). In addition to the semistructured interview and record forms, there is also a self-report form that may be completed by the client which is referred to as the Structured Clinical Interview for DSM-5 Screening Personality Questionnaire (SCID-5-SPQ). The SCID-5-SPQ takes approximately 20 minutes to complete and when used will decrease time spent conducting the semistructured interview. In addition to the SCID-5-PD, First and colleagues (2018) recently released the SCID-5-AMPD, which allows for assessment of personality pathology consistent with the DSM-5 AMPD, and consists of three modules. Module I is the Structured Clinical Interview for the Level of Personality Functioning Scale and allows for assessment of the four domains of functioning (Identity, Self-direction, Empathy, Intimacy) using the Level of Personality Functioning Scale. Module II is the Structured Clinical Interview for Personality Traits and allows for detailed assessment of the five pathological personality trait domains (Negative Affectivity, Detachment, Antagonism, Disinhibition, and Psychoticism) and their trait facets. Module III is the Structured Clinical Interview for Personality Disorders and allows for comprehensive assessment of six Alternative Model personality disorders (Schizotypal, Antisocial, Borderline, Narcissistic, Avoidant, Obsessive—Compulsive) and Personality Disorder—Trait Specified.

Because the personality disorder criteria did not change from the DSM-IV to the DSM-5, these have not been revised. However, the SCID authors thoroughly

revised the SCID-II interview questions to ensure that they optimally capture symptoms necessary to meet diagnostic criteria, provided a dimensional rating system, and added the AMPD modules, making the SCID-5-PD/AMPD a significantly different semistructured interview than the SCID-II. As new measures, the SCID-5-PD/AMPD have limited reliability and validity evidence supporting their use. Their predecessor, the SCID-II, demonstrated good concordance in diagnostic frequency with the Personality Disorder Examination (PDE) in most groups, although some evidence suggest that the SCID-II is overly sensitive to paranoid and schizoid personality disorder diagnoses compared to the PDE (Oldham et al., 1995). Studies on the SCID-II also demonstrated adequate to good interrater reliability for the SCID-II (Lobbestael et al., 2011; Maffei et al., 1997; Weertman, Arntz, Dreessen, Velzen, & Vertommen, 2003). The SCID-II was not as widely adopted as the SCID-I in research or clinical settings. It remains unknown whether or not the SCID-PD will be used more widely, although its careful and comprehensive revision process provide a set of modern rating procedures that are likely to improve diagnostic reliability and advance research in personality disorders.

## Symptom and Behavior Rating Scales

Many rating scales for psychiatric symptoms are currently available. In fact, Rush, First, and Blacker (2008) describe more than 275 different rating scales, which differ based on symptom coverage, length, and clinical expertise required to complete them. Information used to complete rating scales may come from a variety of sources, but a clinical interview with the adult client is almost always the main source of information. The interview format is typically either flexibly structured, semistructured, or structured, because the information covered in the interview is standardized. Various levels of scheduling are used to collect the interview data, depending on the rating scale. Some rating scales provide probe questions to elicit initial information, which are then followed by additional questions from the clinician based on his or her judgment of when additional information is required. Some rating scales require information to be obtained in a prespecified order, while others allow the clinician to move through the content flexibly based on client or interview specific factors. Each scale will specify how to record data from the interview, typically by providing a standardized Likert-type rating for each item with behavioral anchors specified for each rating point. Also, most scales require some degree of training to reliably and validly complete (Ventura, Green, Shaner, & Liberman, 1993).

Some rating scales are designed for assessment of specific psychiatric populations (e.g., depression, PTSD, schizophrenia), and others are multidimensional in that they cover a wide array of symptoms and are designed for use with general psychiatric populations. However, there are numerous common features that guide selection of scales for particular purposes, and distinguish those that are considered

excellent from those that should be avoided. First, there should be a manual included with the rating scale that, at a minimum, explains the purpose of the scale, its development and psychometric properties, and which includes a detailed description of rating procedures. The description of rating procedures should provide: (1) a description of a behaviorally anchored rating scale on which to record the item ratings, (2) the time frame for which the item is to be rated (e.g., today, past week), (3) an operational definition for each item, and (4) the sources of information used to complete the item. More details on these points are included in the following sections.

Ratings are typically recorded on a Likert format scale, with lower ratings indicating the absence of symptoms or mild symptoms, and higher ratings indicating more severe symptoms. Ratings are behaviorally anchored to establish a clear understanding of how to assign a score based on the severity of present symptoms. Behavioral anchors to establish symptom severity are often linked to an objective indicator, such as frequency of the symptom over the past week (occurred once or twice, occurred daily, occurred daily and was continuous), the degree to which the symptom impairs functioning (minimal or no impact on daily function, substantially influences daily behavior), or a combination of the two. In some cases, a symptom that occurs infrequently but has substantial impact on behavior may be rated as severe, as would be the case for an individual who experiences infrequent command hallucinations to harm others but acts in accordance with the commands when they do occur. Behavioral anchors improve reliability of ratings ([Ventura et al., 1993](#)), making scale with clear behavioral anchors preferable.

Importantly, rating scales should include a specified timeframe. Typically, ratings scales will indicate that symptoms are to be rated based on their presence over the past week (including the interview day). However, the timeframe for ratings may be determined by goals of the rating interview and some scales allow the rater to adjust the time frame according to a specific need. Ratings which occur in the days immediately preceding the interview provide a good indicator of current level of symptom severity and are appropriate in many situations; for example, judging week-to-week improvement in symptoms for individuals participating in weekly therapy. Ratings that reflect symptoms over longer time periods (past month) would provide an indication of the more general or characteristic level of symptoms and may be appropriate when, for example, a patient is being followed long-term after successful treatment of an acute depressive or psychotic episode and there is a need to quantify severity of residual symptoms. Although less common, the rating period could also reference a historical period in a client's life, such as the first time depressive or psychotic symptoms were experienced. Generally, the longer the reporting time frame, the less reliable the symptom rating will become due to inaccurate recollection of symptoms. Also, it may be necessary to adjust the behavioral anchors of the rating scale if the time frame is adjusted, particularly when ratings are based on frequency of symptom occurrence.

Operational definitions for each symptom item are critical for reliable assessment of the item because of differences in definitions of common psychiatric

symptoms, rater biases, and other factors. For example, a severity rating of “insomnia” over the past week for an individual complaining of disrupted sleep will be substantially influenced by the number of hours of lost sleep that are used to determine insomnia severity. This essential criterion is quite likely to vary among raters but reliability could be improved if, for example, the scale developers indicate that “severe” insomnia is defined as disrupted sleep every night for the past week with the loss of two or more hours of sleep each night. Scale developers may also provide specific instructions about whether the rater should attribute a symptom rating to an underlying cause. Generally, assumptions or attributions about underlying causes of symptoms are discouraged because they can decrease reliability. For example, symptoms of depression attributed to underlying feelings of anger may lead to a decreased depression severity rating and increased anger rating, or decreased spontaneous movements attributed to side effects of antipsychotic medication may lead to decreased severity rating for blunted affect. At times, such attributions are required (e.g., differentiating between primary and secondary negative symptoms in schizophrenia) and when this is the case, the developer will often provide relevant information that may be considered to improve the reliability of the final rating. Items may be further organized according to symptoms domain (e.g., positive and negative symptoms of schizophrenia) and a severity score may be assigned for the overall domain, in addition to ratings for each symptom within the domain. An example would be a positive symptom rating scale for schizophrenia that requires ratings for various types of hallucinations (auditory, visual, olfactory, tactile, gustatory) and an overall rating for hallucinations as a domain. Some scales may also allow for subjective ratings which are provided by the client to evaluate subjective experience and insight.

Sources of information used to complete each item rating should be clearly indicated, and always include information obtained directly from the interview based on client report (e.g., feelings of sadness), behavioral observation (e.g., diminished facial expression), or medical examination (e.g., cogwheeling identified on motor examination). Other common sources of information are significant other informants or treatment providers and observation of behavior in the hospital (for inpatients) or in other settings (day treatment programs, group residence homes). Although most items will be completed based on data gathered in the interview itself, the source of information may vary from one item to the next. For example, because some items may have lower reliability when client self-report of behavior over the past week is used as the basis for the item rating, the clinician may be directed to rely only on symptoms or behaviors present during the interview. Symptoms like flat affect may be rated based on behavior in the interview, while symptoms of anhedonia might be rated based on clients’ reports of activity level and experience of pleasure over the past week.

In the following sections, we describe a number of ratings scales that are commonly used to assess various psychiatric disorders according to 1) those that are multidimensional and assess a broad range of psychiatric symptoms and 2) those that focus on a particular disorder or specific symptom dimension, emphasizing rating scales for affective and psychotic symptoms.

## **Multidimensional Rating Scales**

### **Brief Psychiatric Rating Scale**

The Brief Psychiatric Rating Scale (BPRS) was originally developed by [Overall and Gorham \(1962\)](#) and continues to be among the most popular behavioral rating scales in use today ([Hafkenscheid, 1993](#)). The authors developed the BPRS to provide a standardized method to assess changes in symptoms in response to medications in outcome studies. However, since its development, it has had widespread application for clinical and research purposes in both inpatient and outpatient settings to assess individuals with a variety of psychiatric disorders. The original BPRS had 16 items representing common psychiatric symptoms such as positive and negative psychosis, mania and depression, anxiety, and somatic disturbances. Since then, important modifications have been made to the original version, including increasing the number of items to increase symptom coverage. To improve reliability and validity of ratings, more detailed operational definitions for items and behavioral anchors to assist in establishing symptom severity were developed. A variety of semistructured interviews were also written to assist with administration and ensure more uniform coverage of information across evaluators ([Bigelow & Murphy, 1978](#); [Lukoff, Ruechterlein, & Ventura, 1986](#); [Overall, Hollister, & Pichot, 1967](#)).

The most recent version of the BPRS is referred to as the BPRS extended version (4.0). It consists of 24 items rated on a seven-point Likert-type scale. A rating of “1” indicates the absence of symptoms, ratings of “2–3” indicate “very mild” to “mild” symptoms that are considered to have nonpathological intensity, and ratings of “6–7” indicate “severe” or “extremely severe” symptoms associated with significant distress or impairment ([Ventura et al., 1993](#)). To improve reliability, guidance is provided regarding the sources of information used to complete each item, with some items based on self-reported symptoms and others based on behavioral observation of the client during the interview. Symptoms are rated for the 2-weeks preceding the assessment. A total score is calculated by summing ratings across all items, but subscales scores are sometimes calculated based on factor analytic studies indicating the BPRS assesses several symptom domains which is expected given the multidimensional nature of the scale. The extended version of the BPRS and supporting materials are contained in [Ventura et al., 1993](#). A child version of the BPRS is also available, the BRPS-C ([Overall & Pfefferbaum, 1982](#); [Shafer, 2013](#)).

Psychometric investigations of various BPRS versions provided evidence for satisfactory to excellent interrater reliability ([Hafkenscheid, 1993](#); [Lachar et al., 2001](#); [Roncone, Ventura, Impallomeni, Fallon, & Morosini, 1999](#); [Ventura et al., 1993](#); [Zanello, Berthoud, Ventura, & Merlo, 2013](#)). There is also evidence for satisfactory validity based on score correlations with other rating scales ([Inch, Crossley, Keegan, & Thorarinson, 1997](#); [Morlan & Tan, 1998](#)) and longitudinal sensitivity to changes in psychiatric symptoms ([Zanello et al., 2013](#)). There has also been substantial interest in the factor structure of the BPRS given that it is a multidimensional rating scale and as a result, interpretation of its total score may be less meaningful than interpretation of subscale or factor scores. Interpretation of

subscales or factors scores were hampered to some degree by divergent results from factor analytic studies that report four, five, and sometimes six factor solutions (e.g., Dingemans, Linszen, Lenior, & Smeets, 1995; Mueser, Curran, & McHugo, 1997; Zanello et al., 2013). The most likely contributors to different solutions reported across studies include item number and content of the BPRS versions, type of factor analysis used, and sample characteristics (Shafer, 2005).

Despite these differences, the original four-factor model suggested by Overall et al. (1967) held reasonably well. This model includes Anxious Depression, Hostile Suspiciousness, Thought Disturbance, and Withdrawal Retardation symptom domains. Also, examination of subscale scores proved more effective than the total score for distinguishing between various psychiatric disorders and for determining changes in specific symptom domains to help inform treatment and judge its effectiveness (Lachar et al., 2001; Long & Brekke, 1999; Nicholson, Chapman, & Neufeld, 1995; Van der Does, Dingemans, Linszen, Nugter, & Scholte, 1995). A meta-analysis conducted by Shafer (2005) of 26 factor analytic studies of the 18-item BPRS identified four core factors, and an additional fifth factor. The five factors and BPRS items that load on them are presented in Table 12.1.

For the four core factors, item loadings were strong on each respective factor with little cross loading. When the fifth factor was added (Activation), the mannerisms and posturing and tension items formed a factor with the excitement item, but the items exhibited cross loadings on other factors, thereby prompting questioning of the validity of the Activation factor. There is less consistency regarding the factor structure of the extended 24-item version of the BPRS. However, it may be said that there is general consistency for the affect, positive symptoms, negative symptoms, and resistance factors for the 18-item BPRS and some confidence may thus be placed in these factor scores. Less confidence is warranted for factors such as Activation, Disorganization, and Somatization, which are identified in some studies but not in others, may not attain simple structure, and at times include items that do not seem to be conceptually related (e.g., see activation factor in Table 12.1). Thus, interpretation of factor scores should focus on the four core factors identified by Shafer (2005) and be qualified based on the factor structure (and study) from which they were derived.

**Table 12.1** Factor structure of the 18-item BPRS from Shafer (2005) meta-analysis

Factor name	BPRS items
Affect <sup>a</sup>	Anxiety, guilt feelings, depressive mood, somatic concern items
Positive symptoms <sup>a</sup>	Unusual though content, conceptual disorganization, hallucinatory behavior, grandiosity
Negative symptoms <sup>a</sup>	Blunted affect, emotional withdrawal, motor retardation, disorientation
Resistance <sup>a</sup>	Hostility, uncooperativeness, suspiciousness
Activation	Excitement, mannerisms and posturing, tension

<sup>a</sup>Core factor identified by Shafer (2005).

## ***Present State Examination***

The PSE was developed in 1967 to provide an objective evaluation of psychiatric symptoms based on a semistructured interview. It is among the most commonly used interviews within the ICD framework and has greater popularity outside of the United States. The PSE incorporates operational definitions of symptoms that are queried in a clinical “cross-examination” using questions found to be acceptable to nearly all interviewees (Wing et al., 1990). Symptom reporting is typically completed for the past month and questions are provided and ordered to elicit symptom information. However, the interviewer is free to ask additional questions and modify question order when needed to obtain supplemental information necessary to clarify symptoms. The PSE organizes symptoms into a distinct syndrome structure whereby, for example, depressive symptoms may be subsumed under several diverse syndromes, which is consistent with its focus on detecting the presence of any major disorder rather than associations between specific symptoms and diagnosis (Rogers, 2001; Wing, Birely, Cooper, Graham, & Isaacs, 1967). The ninth edition of the PSE included 140 symptom items that were scored on a 0–2 scale and translated into approximately 40 different languages (Wing et al., 1990).

The PSE is now in its 10th edition which contains an expanded assessment of psychopathology, with 299 symptom items that are organized into two parts. Part I includes neurotic symptoms, eating disorders, and substance use disorder, while Part II covers psychotic symptoms as well as affect, speech, and behavior that are present during the interview (Wing et al., 1990). Symptom items are rated on a 0–3 scale. Higher scores indicate greater severity of symptoms, including a rating for subclinical symptom elevations, based on the following behavioral anchors: 0 = absent; 1 = present to a minor degree, subclinical degree; 2 = present, clinical level; and 3 = present in a severe form. Behavioral symptoms are also rated using a 0–3 point scale that reflects the frequency of the behavior based on the following behavioral anchors: 0 = did not occur; 1 = occurred but probably uncommon or transitory, 2 = occurred on multiple occasions, 3 = present more or less continuously.

The PSE-10 was integrated as the interview component of the Schedules for Clinical Assessment in Neuropsychiatry (SCAN; WHO, 1994). Wing et al. (1990) indicates that the main goals of the SCAN were to allow for rigorous clinical observation, develop common clinical language standardized across different diagnostic systems, and accumulate clinical knowledge as a result of standardization. In addition to the PSE-10, SCAN also includes the Glossary of Definitions, Item Group Checklist (IGC), and Clinical History Schedule (CHS). The IGC is composed of 59 ratings based on secondary sources such as records or informant. The CHS is an optional section that includes 88 items that focus on childhood, intellectual functioning, social relationships, adult personality, clinical diagnoses, and physical illness. The same 0–3 point rating system from the PSE-10 is used for these scales and there are additional ratings used for attributions regarding etiology and lifelong traits. The SCAN takes 60–90 minutes to complete and provides a highly detailed structured interview containing standardized questions and additional probes that may be used to clarify symptoms. Use of the SCAN and associated computer

programs allows for the data gained from the PSE and other SCAN components to be used for establishing DSM-IV and ICD-10 diagnoses (WHO, 2004). Several studies support moderate to high reliability and validity for the PSE (Huxley, Korer, & Tolley, 1987; Lesage, Cyr, & Toupin, 1991; Mignolli, Faccinican, Turit, Gavioli, & Micciolo, 1988; Peveler & Faiburn, 1990; Wilmink & Snijders, 1989). Psychometric studies of SCAN data gathered from extensive WHO field trials and other sources suggest moderate to high reliability and validity (Brugha et al., 1999; Easton et al., 1997; Farmer et al., 1993, 1996; Hapke, Rumpf, & John, 1998; Wing et al., 1990, 1998).

### *Positive and Negative Syndrome Scale*

The Positive and Negative Syndrome Scale (PANSS; Kay, Fiszbein, & Opler, 1987) was developed to evaluate symptoms in individuals with schizophrenia, but is reviewed here because of its coverage of other symptoms in addition to psychosis. It is among the most popular rating scales for schizophrenia and although it is used in clinical settings, it has had broad application in clinical pharmaceutical trials. The PANSS contains 30 items rated on 1 (absent) to 7 (extreme) scales. Items are divided into three symptoms domains that include positive symptoms (7 items), negative symptoms (7 items) and general psychopathology (16 items). The psychopathology symptom domain includes symptoms such as depression, anxiety, motor retardation, social avoidance, and poor attention. Scores can be derived for each symptom domain and a total score can also be calculated. Symptoms are typically rated for the week immediately preceding the interview, which takes 30–40 minutes to complete. Operational definitions and behavioral anchors are provided to increase reliability and a manual is available with probe questions to provide standardization and scheduling for the interview. Symptoms ratings are based primarily on client report and behavior observed during the interview, but information from collateral sources such as mental health providers, family members, or others who are familiar with the client may also be used. Items for the PANSS were selected from the Psychopathology Rating Schedule (Singh & Kay, 1987) and BPRS, thus it has considerable overlap with those scales. A child version of the PANSS is also available (Fields et al., 1994).

Studies of the psychometric properties of the PANSS suggest that the total and domain scores have adequate reliability and validity (see Lindenmayer, Harvey, Khan, & Kirkpatrick, 2007; for review). Factor analytic studies suggest that the PANSS is probably best understood as assessing five symptom domains rather than three, including positive symptoms, negative symptoms, cognitive or disorganized symptoms, depression, and excitement (Rodriguez-Jimenez, Bagney, Mezquita, Martinez-Grasa, & Sanchez-Morla, 2013). The factor structure of the PANSS and BPRS are similar in many respects, but for the PANSS it is important to keep in mind when interpreting scores that the factors are composed of items that span the three originally proposed symptoms domains, for example, the disorganized/cognitive factor is composed of three items from the positive (conceptual disorganization), negative (difficulty in abstract thinking), and general (Poor Attention)

domains. This finding suggests that interpreting scores according to the original positive, negative, and general domains is questionable, although interpretation of factor scores is not commonplace. As with the BPRS, if factor scores are used for clinical or research purposes, it would be important to identify the study from which the scores were derived.

### ***Symptom specific and clinical diagnosis rating scales***

#### ***Affective symptoms***

Affective disorders which are characterized by the presence of depressive and manic symptoms are common psychiatric disorders, particularly major depressive disorder, with an estimated 16.2 million or 6.7% of adults in the United States experiencing a major depressive episode in 2016. Of these, 64% or 10.3 million US adults experienced severe impairment as a result ([SAMHSA, 2017](#)). Although less common, an estimated 2.6% of US adults have bipolar disorder which causes significant disability in 83% of affected individuals ([Alegria et al., internet citation, retrieved 6/1/18](#)). Symptoms of depression and mania also occur at a high rate in nonaffective disorders. For example, prevalence rates of major depressive episodes are as high as 61% in schizophrenia ([Gozdzik-Zelazny, Borecki, & Pokorski, 2011](#)), 50% in anxiety disorders, and 33% in substance use disorders ([Davis, Uezato, Newell, & Frazier, 2008](#)). Therefore evaluation of affective symptoms and disorders are an essential component of most clinical interview procedures and many rating scales were developed to assist in this purpose.

Development of rating scales for depression and mania typically progressed by developing separate scales. The symptom coverage for each scale varies particularly from older to more newly developed scales which in part is due to the ongoing development of more precise diagnostic criteria and changes in conceptual understanding of affective disorders. Structure of manic and depressed symptoms have been studied in a different ways, including factor analysis of rating scales designed to assess these symptoms. For depressive symptoms, factor analytic studies are limited to some degree based on the assessment procedure used to evaluate depression, which vary in terms of symptom coverage. A recent analysis of three depression measures and an anhedonia scale in a sample of 119 inpatients with unipolar or bipolar depression indicated the presence of eight factors including depressed mood, tension, negative cognition, impaired sleep, suicidal thoughts, reduced appetite, anhedonia, and amotivation ([Ballard et al., 2018](#)). Examination of self-report scales such as the Beck Depression Inventory suggest the presence of a general depression factor and a somatic and cognitive factor ([Quilty, Zhang, & Bagby, 2010; Ward, 2006](#)). The DSM-IV criteria load on two factors, one “somatic” (sleep disturbance, fatigue/lack of energy, eating disturbance, trouble concentrating, and psychomotor disturbance), and the other “psychological” factor (depressed mood, lack of interest/pleasure, worthlessness/guilt, and suicidal thoughts) ([Sunderland, Carragher, Wong, & Andrews, 2013](#)). The studies demonstrate the importance of symptom coverage of the scale used to assess depression, with scales that provide

broader and more extensive coverage allowing for distinctions between a relatively large number of symptoms.

As in depression, mania is also viewed as a multicomponent construct. As early as the third century, Aretaeus wrote that some individuals with mania are “cheerful, they laugh, play, dance day and night” while others “fly into a rage” (Marneros, & Angst, 2002). Robertson (1890) made similar observations, suggesting two types of mania described as “hilarious” and “furious.” Review of early factor analytic studies suggests that in addition to euphoric mood and irritable aggression, two other classic features defining mania are psychomotor pressure and grandiosity (Cassidy, Forest, Murry, & Carroll, 1998; Cassidy, Murry, Forest, & Carroll, 1998). A more recent review of 23 studies examining the factor structure of mania published between 1970 and 2016 noted a median of five symptom dimensions among which activation, elation, and depression or dysphoria emerge as the most common primary factors across studies (Scott et al., 2017). Other common factors included irritability/hostility/aggression, psychosis, and sleep disturbance. For mixed episodes, the depression/dysphoria factor emerged most consistently as primary, and for pure manic episodes activation was the most common primary factor. Identification of activation as a primary factor in manic episodes supports the DSM-5’s inclusion of increased activity or energy as a symptom required for mania and with others who have proposed that increased psychomotor activity is the most consistent and prominent feature of bipolar mania (Alloy & Abramson, 2010; Johnson, Edge, Holmes, & Carver, 2012). These mania factors are most readily identified in patients experiencing acute manic or mixed episodes (Harvey, Endicott, & Loebel, 2008).

Notably, the DSM used a system of defining manic or depressed episodes to arrive at various affective disorder diagnoses. Manic and depressive symptoms may occur together which was why until the DSM-5, there were also criteria available to diagnose a mixed episode. Kraepelin (1921) described this phenomenon as anxious or depressed mania, and in clinical settings, it is often noted that when patients experience manic episodes, dysphoria or other symptoms of depression are also present. Dilsaver, Chen, Shoaib, and Swann (1999) identified three types of manic episodes by cluster analyzing four symptom factors from the Schedule for Affective Disorders and Schizophrenia (SADS) (euphoric mania, sleep disturbance, depression, irritability/psychosis) in a sample of 105 individuals hospitalized for treatment of mania. The most common type of episodes was referred to as pure or euphoric mania which included 45 of the patients. However, also identified were dysphoric manic episodes ( $n = 30$ ), and depressed manic episodes ( $n = 27$ ). Euphoric mania was characterized by grandiosity and euphoria with minimal depressive symptoms, while depressed mania was characterized by high depressive symptoms and elevated sleep disturbance, manic euphoria, and irritability. Dysphoric manic episodes included both manic and depressed symptoms that were intermediate to the other two-episode types and elevations on irritability and sleep disturbance. These and other results (e.g., Cassidy, Forest, 1998; Cassidy, Murry, 1998; Swann et al., 2013) indicate that manic and depressive symptoms are not mutually exclusive phenomena, and both should be considered in the assessment of manic episodes.

Many clinician-rated and self-report assessment procedures for depression exist, including the Hamilton Rating Scale for Depression (HRSD), which is among the oldest and most widely used, the Inventory of Depressive Symptomatology (IDS), which is a newer rating scale designed to address some of the shortcomings of older ratings scales, and the Calgary Depression Scale for Schizophrenia (CDSS), which is a specialized rating scale for assessing depression in schizophrenia. Our reviews do not consider differences in expression of depression in late life (Hegeman et al., 2012) although there are assessment procedures designed for this purpose including the Geriatric Depression Scale (Yesavage, 1988) and the Cornell Scale for depression in Dementia (Alexopoulos, Abrams, Young, & Shamoian, 1988). For mania, we review the Young Mania Rating Scale (YMRS; Young, Biggs, Ziegler, & Meyer, 1978) and the Beck–Rafaelsen Scale (Bech, Bolwig, Kramp, & Rafaelsen, 1979) because these two scales are the most widely used in clinical trials, have extensive psychometric support, and provide adequate coverage of the major symptoms of mania. Information regarding late life changes in symptoms of mania is limited, but the YMRS is commonly used in clinical trials of older individuals (Young, Peasley-Miklus, & Shulberg, 2007). It has also been validated in pediatric bipolar disorder (Youngstrom, Danielson, Findling, Gracious, & Calabrese, 2002).

### Hamilton Rating Scale for Depression

The HRSD (Hamilton, 1960) is among the oldest and most widely used depression rating scale. Its original version consisted of 21 items that assessed depressive and other symptoms. The more recent 17-item version eliminated items that were not directly associated with depression (e.g., depersonalization, paranoid, and obsessive–compulsive symptoms) with the retained items assessing symptoms more directly related to depression including depressive thoughts and feelings, anhedonia, suicidal ideation, somatic concerns, weight changes, insomnia, and anxiety. Items are rated on scales that range from 0–2, 0–3, or 0–4, depending on the item, with 0 indicating the absence of symptoms and higher scores indicating progressive increases in symptom severity. Behavioral anchors are provided for each rating but unlike other ratings scales, the convention of mild, moderate, severe, is not typically used. Rather, a descriptor is provided for each item, for example, selected ratings for the suicide item include: “0, Absent; 2, Wishes he were dead or any thoughts of possible death to self; 4, Attempts at suicide (and serious attempt rates 4).” Definitions are not provided for each item, although the descriptive behavioral anchors help address concerns in this regard. Ratings are made based on an interview with the client which can be completed in 15 minutes. While the original publication did not include a semistructured interview, a number were published since (e.g., Williams, 1988; Williams et al., 2008), which proved helpful for improving interrater reliability. Ratings are typically completed for the week prior to the interview.

Total scores for the HRSD range between 0 and 54, with scores of 0–6 indicating the absence of depression, 7–17 indicating mild depression, 18–24 indicating moderate depression, and scores of 25 or higher indicating severe depression. Reliability studies suggest that use of the structured interview improves reliability

(Hamilton, 2000), with internal consistency estimates of 0.83 (Rush et al., 2003), total score interrater reliability of 0.80–0.98 (Moberg et al., 2001), and test–retest reliability as high as 0.81 (Davidson, Turnbull, Strickland, Miller, & Graves, 1986; Takahashi, Tomita, Higuchi, & Inada, 2004; Williams, 1988). Validity evidence for the HRSD is also apparent based on high correlations between the HRSD total score and other depression ratings scales (Hamilton, 2000). While a total score in the most commonly used, factor analytic studies indicate that the HRSD is a multi-dimensional measure, although the structure of the scale remains debated as studies found anywhere from 3 to 6 factors (e.g., Brown, Schulberg, & Madonia, 1995; Marcos & Salamero, 1990). A recent meta-analysis that examined 17 factor analysis studies ( $N = 2606$ ) published from 1967 to 2001 indicated that the 17-item version is composed of four factors termed anxiety, depression, insomnia, and somatic (Shafer, 2006). Limited information is available about the usefulness of these factors in clinical or research settings (<http://serene.me.uk/tests/ham-d.pdf>).

### Inventory of Depressive Symptomatology

The IDS (Rush et al., 1986; Rush, Gullion, Basco, Jarrett, & Trivedi, 1996) was developed to address shortcomings in older widely used depression rating scales like the HRSD and the Montgomery Asberg Depression Rating Scale (MADRS; Montgomery & Asberg, 1979), including limited symptom coverage and measurement issues related to item scaling and weighting. The IDS expands symptom coverage to include items reflecting all depression symptoms from the DSM and melancholic and atypical symptoms. Measurement issues were addressed by implementing a uniform rating scale allowing for equivalent weighting of each item and providing clear behavioral anchors for each item rating that reflect severity and frequency of depressive symptoms. There are two versions of the IDS, one that contains 30 items (Rush et al., 1996) and another that contains 16 items and is named the Quick Inventory of Depressive Symptomatology (QIDS; Rush et al., 2003). The QIDS was developed to provide a more efficient means of assessing depression in research and clinical settings by selecting IDS items to reflect each of the nine DSM symptom domains. Client self-report versions that are matched to the IDS and QIDS are also available (Rush et al., 1986, 1996, 2003). For the clinician-rated (IDS-C, QIDS-C) and client self-report (IDS-SR, QIDS-SR) versions, the rating period is the week prior to the evaluation. Scores for the IDS-C and IDS-SR range from 0 to 84 and scores for the QIDS-C and QIDS-SR range from 0 to 27. Cutoff scores for depression severity are provided on the website for each scale. For the IDS-C the following cutoff scores are used: None = 0–11; Mild = 12–23; Moderate = 24–36; Severe = 37–46; Very Severe = 47–84. The IDS-C takes 10–15 minutes to administer and the QIDS-C takes 5–7 minutes. A semistructured interview with probe questions is available for the IDS-C and QIDS-C which helps standardize administration although the IDS authors suggest that experienced clinicians will be able to reliably complete the IDS without the use of the manual. The IDS has been used to assess individuals in which depressive symptoms are a component of the clinical presentation including major depressive disorder, bipolar disorder, and dysthymia, among others.

The authors provide extensive reliability and validity evidence supporting the IDS on their website ([www.ids-qids.org](http://www.ids-qids.org)). For example, Trivedi and colleagues (2004) examined psychometric properties of all IDS scales in sample of individuals with major depressive disorder ( $n = 544$ ) and bipolar disorder ( $n = 402$ ). Internal consistency estimates for all four scales were high and ranged from 0.81 to 0.92. Interrater reliability is also high for the IDS-C (0.96). Evidence for concurrent validity of IDS scales was provided by high correlations of total scores ( $r = 0.81\text{--}0.83$ ) and equal sensitivity to symptom change. There are also strong correlations ( $r = 0.61\text{--}0.92$ ) between IDS and QIDS total scores and other depression rating scales like the HRSD and Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). These reliability and validity estimates compare quite favorably to other commonly used interview-based depression measures such as the HRSD and MADRS. The QIDS has comparable reliability and validity data to the IDS (Rush et al., 2003; Trivedi et al., 2004). The factor structure of the IDS was also explored and results suggest that it is multidimensional, assessing three depression dimensions described as cognitive/mood, anxiety/arousal, and vegetative (Rush et al., 1996; Wardenaar et al., 2010) although the factor structure appears sensitive to age effects, as mood, motivation, and somatic symptoms factors are identified in older adults (Hegeman et al., 2012). More information about the IDS in its various forms, including the administration and scoring procedures, interpretation guidelines, and test forms is available at [www.ids-qids.org](http://www.ids-qids.org).

### Calgary Depression Scale for Schizophrenia

The CDSS (Addington, Addington, & Schissel, 1990; Addington, Addington, & Maticka-Tyndale, 1993) was specifically designed to assess severity of depression in individuals with schizophrenia and is currently the most frequently used scale for that purpose (Rabany, 2013). The CDSS is more accurate in identifying depressive symptoms in schizophrenia compared to the other major rating scales such as the HRSD (Schennach et al., 2012) because it was designed to minimize overlap between symptoms of schizophrenia and those that occur in depression. For example, anhedonia may be present in depression and psychotic disorders and some anti-psychotic medications produce blunted affect which may be confused as a symptom of depression even though it is a medication side effect.

As a result, strong correlations are reported between the CDSS total score and other depression rating scale scores, while correlations with measures of negative symptoms and extrapyramidal side effects are minimized (Addington, Addington, & Maticka-Tyndale, 1994; Lako et al., 2012). The CDSS consists of nine items reflecting various aspects of depression (depression, hopelessness, self-deprecation, guilty ideas of reference, pathological guilt, morning depression, early wakening, suicide, and observed depression) that are rated over the past 2 weeks. Each item is rated on a 0–3 point scale, with “0” indicating the absence of symptoms and “3” indicating severe symptoms. Behavioral anchors are provided to make these ratings. Probe questions are provided which elicit information needed to make all ratings except for the “observed depression” rating, which is based on observation during the interview. Psychometric studies suggest adequate internal consistency estimates

that ranges from 0.76 to 0.82, test–retest reliability that range from 0.69 to 0.93, and an interrater reliability of 0.86 (Addington et al., 1994; Lako et al., 2012). The CDSS total score correlated highly with scores from other depression measures ( $r = 0.66$ – $0.81$ ) and classification accuracy is good, averaging 88% correct classification across studies although cutoff scores vary among studies probably due to differences in sample characteristics. Factor analytic studies suggest that the CDSS may be best characterized by two or three factors reflecting depression, guilt, and in some analyses morning depression (Maggini & Raballo, 2006; Rabany, Weiser, & Levkovitz, 2013; Rekhi, Ng, & Lee, 2018; Schennach et al., 2012), although there is limited information regarding the utility of factors for research and clinical purposes.

### Young Mania Rating Scale

The YMRS (Young et al., 1978; Young, Biggs, Ziegler, & Meyer, 2000) is by far the most widely used rating scale for manic symptoms. It is composed of 11 items that assess elevated mood, activity-energy, sleep, sexual interest, speech, irritability, thought disorder, abnormal mental content, aggressive behavior, appearance, and insight. These items were selected to represent the core symptoms of mania. Seven items are rated on a 0–4 scale and the other four items are double weighted using a 0–8 scale (irritability, speech, thought content, and disruptive/aggressive behavior) to account for poor cooperation from severely ill individuals. Half point ratings may be assigned in cases when severity of the symptom falls between two anchors. Ratings are based on client self-report of mood and behavior over the previous 48 hours and behavioral observations made during a clinical interview which takes 15–30 minutes to complete. No semistructured interview with probe questions is provided to assist with standardization of the clinical interview. Total scores range from 0 to 60 with higher scores indicating more severe manic symptoms.

Definitions for items are not provided, but descriptive behavioral anchors are included. Examples of anchors for the “insight” item include—0: Present, admits illness, agrees with need for treatment; 2: Admits behavior change, but denies illness; 4: Denies any behavior change. For reliability and validity, the authors report interrater reliability of 0.93. Internal consistency estimates reported in child samples range from 0.80 to 0.91 (Youngstrom et al., 2002; Fristad, Weller, & Weller, 1995). The YMRS authors provide evidence for convergent validity demonstrated through correlations with independent ratings on other mania scales for adults ( $r$ 's = 0.71, 0.89; Young et al., 1978), and for children ( $r = 0.83$ ; Fristad et al., 1995). Reliability and validity also appear good for some translated versions of the YMRS, with a study of the Spanish translation reporting internal consistency of 0.88, test–retest reliability of 0.76, and good convergent validity when compared to another mania rating scale (Colom et al., 2002). Clinical trials also provide ample evidence that the YMRS is sensitive to changes in symptoms of mania resulting from treatment.

Understanding of the factor structure of the YMRS and other mania scales is limited because they have typically been examined with measures of depression, which substantially influences the identified structure. In the first factor analysis of the YMRS, Beigel and Murphy (1971) identified two factors they named

“paranoid-destructive” and “euphoric-grandiose” but only 12 patients were included in that study. Hanwella and de Silva (2011) extracted three factors in a sample of 131 patients referred to as “elated mania” (elated mood, language abnormalities/thought disorder, increased sexual interest, poor insight), “irritable mania” (irritability, increased motor activity/energy, disruptive aggressive behavior), and “psychotic mania” (abnormal thought content, impaired self-care, poor sleep, speech abnormalities). Since there are no items assessing depressed mood on the YMRS, a depression/dysphoria factor was not identified in their analysis.

### Bech–Rafaelsen Mania Scale

The Bech–Rafaelsen Mania Scale (MAS) was originally published in 1979 around the same time as the YMRS (Bech et al., 1979). The scale was intended to complement existing rating scales developed by Hamilton for the assessment of depression and anxiety (Hamilton, 1959, 1960) who had not developed a comparable scale for rating mania. It is widely used in treatment and basic research second only to the YMRS (e.g., Bech, 2002; Johnson et al., 2008; Malkoff-Schwartz et al., 1998). The MAS consists of 11 items that assess motor activity, verbal activity, flight of thoughts, voice/noise level, hostility/destructiveness, mood (feelings of wellbeing), self-esteem, contact with others, sleep changes, sexual interest, and work activities. Each item is rated on a five-point scale with 0 indicating normal mood and behavior and “4” indicating severe impairment. The total score can be used to reflect the severity of mania as either mild (15–20), moderate (21–28), marked (29–32), severe (33–43), and extreme ( $\geq 44$ ) (Bech, Bastrup, de Bleeker, & Ropert, 2001). Behavioral anchors are provided for each rating. Ratings are made for symptoms that occur over the past three days based on a clinical interview with the client which takes between 15 and 30 minutes to complete.

Internal consistency and interrater reliability are both excellent with interrater reliabilities ranging between 0.89 and 0.99 and an internal consistency estimate for the total score of 0.90 (Bech, 2002; Bech et al., 1979). Regarding validity, the MAS author suggests the scale has good content validity as it provides adequate coverage of the DSM mania symptoms and is comparable to the YMRS in this respect (Bech, 2002). The MAS is sensitive to effects of treatment of manic symptoms (Bech, 2002), and significant correlations are present between MAS endpoint remission scores and blood concentrations of haloperidol ( $r = -0.71$ ) (Gjerris et al., 1980) and carbamazepine ( $r = -0.61$ ) (Chen, 1990). The MAS is equally responsive to reduction in symptoms compared to the YMRS and Clinician-Administered Rating Scale for Mania (Shansis, Reche, & Capp, 2016). Few factor analyses of the MAS have been conducted, but their conclusions seem to indicate that it measures one underlying construct (Bech et al., 2001; Licht & Jensen, 1997; Rossi et al., 2001).

### *Psychotic symptoms*

Before discussing specific rating scales for psychotic symptoms, some background information about the structure of psychotic symptoms is useful. Although traditionally divided according to positive and negative symptom dimensions, research has

demonstrated that the structure of symptom dimensions in psychotic disorders is actually quite complex. A three-dimensional model that includes positive, negative, and disorganized symptoms received substantial support (Andreasen, Arndt, Alliger, Miller, & Flaum, 1995), and from a clinical perspective, provides a picture of current severity and profiles along these three important dimensions. However, it is also understood that these three dimensions might be higher-order constructs, each of which is composed of lower order symptom dimensions which are necessary to fully characterize the complexity of psychotic symptoms (Peralta & Cuesta, 1999). Examples of positive symptoms include hallucinations and delusions. Affective flattening, alogia, avolition and apathy, and anhedonia and asociality are considered negative symptoms. Disorganization symptoms include both disorganized thinking (thought disorder) and bizarre or disorganized behavior, and some studies suggest that attention impairment and alogia also load on a disorganization factor (Bilder, Mukherjee, & Rieder, 1985). Considering that formal thought disorder is the core feature of disorganization, it might be best conceptualized as a positive symptom factor. Disorganized symptoms are the most temporally unstable of the three symptom dimensions (Peralta & Cuesta, 2001) and appear to fluctuate when treatment is effective and there is associated remission of thought disorder. Positive symptoms such as hallucinations and delusions are also temporally unstable and often diminish in response to effective treatment.

On the other hand, negative symptoms tend to be more stable and frequently do not respond to first- or second-generation antipsychotic medications (Carpenter, Conley, Buchanan, Breier, & Tamminga, 1995) leading to high incidence of residual negative symptoms (Carpenter, Heinrichs, & Wagman, 1988). Negative symptoms are associated with poorer outcomes including poorer functioning, greater psychopathology, and higher treatment noncompliance (Galderisi et al., 2013). A distinction is made between negative symptoms which are primary and those that are secondary (Carpenter et al., 1988; Kelley, Gilbertson, Mouton, & van Kammen, 1992). Primary negative symptoms are stable core features of the psychotic disorder and so are not influenced by secondary factors and persist during periods of positive symptom remission. Secondary negative symptoms are not core features of psychotic disorders but rather, are caused by secondary influences as in the case of antipsychotic medication causing affective flattening, depression causing anhedonia, or paranoia causing asociality (Kirkpatrick, Buchanan, McKenney, Alphs, & Carpenter, 1989). Researchers recently distinguished between two negative symptom dimensions reflecting motivation and pleasure (MAP) (anhedonia, avolition, asociality) and diminished expressivity (blunted affect, alogia), with five symptom dimensions that compose these two domains (Blanchard & Cohen, 2006; Kelley et al., 1992; Kirkpatrick et al., 2011; Strauss et al., 2012). Recent findings suggest that conceptualizing the latent structure of negative symptoms as two distinct dimensions does not adequately capture the complexity of negative symptoms, which are better conceptualized along the five distinct symptom domains of anhedonia, avolition, asociality, blunted affect, and alogia (Ahmed et al., in press; Strauss et al., 2018). Whether profile differences across these five domains will have clinical and research application remains to be seen, but there is now abundant

evidence that highlights the importance of assessing negative symptoms in schizophrenia and other psychotic disorders.

In the following sections, we provide a summary of some of the available measures to assess psychotic symptoms. The PANSS, which was previously reviewed, can be used for this purpose. For positive symptoms of psychosis, we focus our discussion on the Scale for the Assessment of Positive Symptoms (SAPS) (Andreasen, 1984) because it provides comprehensive coverage including disorganization symptoms. For negative symptoms, the Scale for the Assessment of Negative Symptoms (SANS; Andreasen, 1983) and the Brief Negative Symptom Scale (BNSS) (Kirkpatrick et al., 2011) are reviewed.

### Scale for the Assessment of Positive Symptoms

The SAPS (Andreasen, 1984) contains 34 items that are organized into four positive symptom domains (hallucinations, delusions, bizarre behavior, and positive formal thought disorder). They are rated on a 0–5 Likert scale with “0” indicating no symptoms and “5” indicating severe symptoms. Ratings are made following an interview conducted by the clinician and based on the client’s subjective reports of behavior and experience, as well as behavioral observations made by the clinician during the interview. The SAPS was originally developed to rate symptoms over the prior month, but the time frame may be adjusted to suit research or clinical demands. The SAPS manual provides definitions of each of the four symptom domains and each of the symptoms to be rated within each domain. Examples are also provided for each symptom along with probe questions and anchor points for the 6-point rating scale.

For example, the hallucinations domain contains six items that assess auditory hallucinations, voices commenting, voices conversing, somatic or tactile hallucinations, olfactory hallucinations, and visual hallucinations. A global severity item is also included. Hallucinations are defined as: “Hallucinations represent an abnormality in perception. They are false perceptions occurring in the absence of some identifiable external stimulus. They may be experienced in any of the sensory modalities . . . True hallucinations should be distinguished from illusions . . . hypnagogic and hypnopompic experiences . . . or normal thought processes that are exceptionally vivid. If the hallucinations have a religious quality, then they should be judged within the context of what is normal for the patient’s social and cultural background” (Andreasen, 1984). The Somatic or Tactile Hallucinations item is defined as “These hallucinations involve experiencing peculiar physical sensations in the body. These include burning sensations, tingling and perceptions that the body has changed in shape or size.” Rating anchor points to determine severity of symptoms include—0, None; 1, Questionable; 2, Mild: has occurred once or twice; 3, Moderate: Occurs at least weekly; 4, Marked—occurs frequently; and 5, Severe: Occurs almost daily. Probe questions include “Have you ever had burning sensations or other strange feelings in your body? What were they? Did your body ever appear to change in shape or size?”

Global ratings are made for each of the four domains. The global ratings do not simply represent a sum of the scores from individual items but account for the

duration and severity of symptoms, their impact on behavior, and other considerations. An auditory hallucination that occurs daily, causes violent and assaultive behavior, and is very distressing for the client may warrant a global score of severe even in the absence of the other five types of hallucinations that are rated in the hallucinations section.

Clinicians can calculate a total score and scores for the four domains. Interrater reliability of the summary score for the SAPS is good ( $r = 0.84$ ; [Norman, Malla, Cortese, & Diaz, 1996](#)) and the summary score is highly correlated with the positive symptom scores from other rating scales such as the PANSS ( $r = 0.91$ ; [Norman et al., 1996](#)).

### Scale for the Assessment of Negative Symptoms

The SANS ([Andreasen, 1983](#)) consists of 30 items and was designed to be used in conjunction with the SAPS to provide a standardized way to evaluate the negative symptoms in schizophrenia. The SANS items are organized to assess five symptom domains including affective flattening or blunting, alogia, avolition—apathy, anhedonia—asociality, and attention. A global severity item is included for each domain. The SANS is structured in a similar manner to the SAPS in that ratings are made based on a clinical interview with the client, symptoms are evaluated over the past month, definitions are provided for each symptom domain and item, items are rated on a 0–5 point scale, examples are provided for specific symptoms, and probe questions are provided when appropriate.

For example, for the anhedonia—asociality domain, anhedonia—asociality is defined as: “This symptom complex encompasses the schizophrenic patient’s difficulties in experiencing interest or pleasure. It may express itself as a loss of interest in pleasurable activities, an inability to experience pleasure when participating in activities normally considered pleasurable, or a lack of involvement in social relationships of various kinds.” It contains six items which assess recreational interests and activities, sexual interest and activity, ability to feel intimacy and closeness, relationships with friends and peers, subjective awareness of anhedonia—asociality, and a global rating of anhedonia—asociality. The “relationships with friends and peers” item is defined as: “Patients may also be relatively restricted in their relationships with friends and peers of either sex. They may have few or no friends, make little or no effort to develop such relationships, and choose to spend all or most of their time alone.” Behavioral anchors to assign ratings include: 0, No inability to form friendships; 1, questionable inability to form friendships; 2, mild but definite inability to form friendships; 3, moderate inability to form friendships; 4, marked inability to form friendships; 5, severe inability to form friendships.

A SANS total score and the five global ratings scores can be calculated. Interrater reliability of the summary scores for the SANS were moderate to good, ranging from 0.60 to 0.84 ([Andreasen & Olsen, 1982](#); [Norman et al., 1996](#)) and the SANS total score is highly correlated with the negative symptom subscale of the PANSS ( $r = 0.88$ ).

### Brief Negative Symptom Scale

The BNSS (Kirkpatrick et al., 2011) is one of numerous second-generation negative symptom scales that were developed as part of the NIMH-sponsored Consensus Development Conference on Negative Symptoms (Kirkpatrick, Fenton, Carpenter, & Marder, 2006). The BNSS was developed to reflect more recent understandings of the negative symptom complex compared to predecessor ratings scales like the SANS. The BNSS contains 13 items that are organized into six symptom domain subscales to assess anhedonia, distress, asociality, avolition, blunted affect, and alogia. Items are rated on a 0–6 point scale, with a rating of 0 indicating normal or no impairment and a rating of 6 indicating presence of an extremely severe symptom or severe deficit.

The BNSS comes with a manual, workbook, and score sheet. Testing materials can be obtained at <http://schizophreniabulletin.oxfordjournals.org/content/37/2/300/suppl/DC1>. The BNSS manual provides definitions for each of the symptom domains, detailed behavioral anchors to rate each item, and probe questions for each item. The probe questions are organized into a semistructured interview that is provided in the BNSS workbook, which is used to obtain information to complete the ratings. Ratings are based on client self-report for symptoms and behaviors over the past week and behavioral observations made during the interview. Information from informants familiar with the client's activities for the rating period may also be considered.

For example, the distress domain consists of one item that is defined as: "This item rates the subject's experience of unpleasant or distressing emotion of any kind: sadness, depression, anxiety, grief, anger, etc. The source of the distress is not considered; for instance, unpleasant emotions associated with psychotic symptoms are considered here." Probe questions for the item in the semistructured interview include "What made you feel bad in the last week? Did anything happen that you didn't like? Did anything make you feel sad or depressed? Worried or anxious? Angry or irritated?" Examples of behavioral anchors to rate the item include: "0, Normal: Normal ability to experience distress and unpleasant emotions; 3, Moderate: Definitely less upset than normal in the face of upsetting events, but does experience some distress; 6, Extremely severe: No experience of distress, no matter what problem is encountered." A total score can be calculated as can subscale scores.

The BNSS has excellent reliability and precision, with an interrater reliability for the total score of 0.96, an internal consistency estimate of 0.93, and a test–retest reliability of 0.81 (Kirkpatrick et al., 2011). Validity information indicates good discriminant, convergent, and predictive validity. Factor analyses identifies either two constructs, one referred to as MAP, and another designated emotional expressivity (EXP; Kirkpatrick et al., 2011; Strauss, Hong, Keller, Buchanan, & Gold, 2012), or five constructs reflecting anhedonia, asociality, avolition, blunted affect, and alogia (Ahmed et al., in press; Strauss et al., 2018). Hierarchical confirmatory factor analysis provide some evidence that MAP and EXP are second-order factors and the five symptom domains are first-order factors, with anhedonia, avolition, and asociality loading on the MAP second-order factor, and blunted affect and alogia loading

on the EXP second-order factor (Strauss et al., 2018). Considered together, interpretation of the five first-order factors is preferable because it likely provides a more detailed understanding of negative symptoms.

The primary goal in development of the BNSS was to construct a scale grounded in current theory that could be used in multicenter trials and completed in approximately 15 minutes. Although limited information is available regarding its use in clinical settings, because it is concise and user-friendly, has detailed testing materials, and excellent psychometric properties, the BNSS may prove to be a valuable clinical assessment procedure.

## Conclusion

We presented a broad range of assessment procedures including the traditional diagnostic assessments, but also more specific rating scales, such as the BPRS, PANSS, HRSD, YMRS, SANS, and SAPS. Discussion focused on multidimensional scales and those designed for rating affective and psychotic symptoms, but there are other well-validated scales for assessment of other symptoms domains that we did not cover. For example, the Hamilton Anxiety Rating Scale (Hamilton, 1960) is widely used to assess the psychic and somatic symptoms of anxiety, as are the Clinician-Administered PTSD Scale (Blake et al., 1995) and the PTSD Checklist (Blanchard, Jones-Alexander, Buckley, & Forneris, 1996) for gathering diagnostic and symptom severity information for PTSD, and the Liebowitz Social Anxiety Scale (Liebowitz, 1987) and Brief Social Phobia Scale (Davidson et al., 1997) for assessment of social phobia. There are the Autism Diagnostic Interview—Revised (Rutter et al., 2003) and Autism Diagnostic Observation Schedule (Lord et al., 2000) for diagnosis and assessment of autism spectrum disorders, and the Eating Disorders Examination (Fairburn & Cooper, 1993) and Eating Disorder Diagnostic Interview (Presnell & Stice, 2003) for eating disorder assessment. Individuals with substance use disorders are often evaluated with the Addiction Severity Index (McLellan, Kushner, & Metzger, 1992) to determine substance use severity and associated problems, while the Timeline Follow Back Assessment procedure (Sobell, Sobell, Klajner, Pavan, & Basian, 1986) can be used to examine daily frequency of drug use over 1–4 months prior to assessment. Rush et al. (2008) provide a comprehensive description of many other commonly used rating procedures, and interested readers are referred there for more information. Clinicians may use these to improve diagnosis and reliability of clinical observations in the initial intake evaluation when the disorder first emerges to determine diagnosis and severity of symptoms, for periodic monitoring with various scales associated with specific symptoms of the disorders, and for monitoring of medication and other treatment and management procedures designed to minimize psychopathology and prevent relapse.

There are other important areas that are often the focus of evaluation which are not covered in this chapter. Some are commonly included in clinical assessments by psychologists depending on the assessment and have direct clinical relevance for

mental disorders. Examples include scales used to assess functional outcomes, such as those that assess activities of daily living (Birchwood, Smith, Cockrane, Wetton, & Copestake, 1990), instrumental activities of daily living (Birchwood et al., 1990; Jaeger, Berns, & Czobor, 2003; Mausbach, Harvey, Goldman, Jeste, & Patterson, 2007; Patterson, Goldman, McKibbin, Hughs, & Jeste, 2001; Schneider, & Struening, 1983), and quality of life (Diamond & Becker 1999; Heinrichs, Hanlon, & Carpenter, 1984). These assessment procedures have good psychometric properties, and although many were developed for research application, they may also have direct clinical applications. Other examples are measures of motor abnormalities, such as extrapyramidal symptoms (dyskinesias), Parkinsonian symptoms, and neurological soft signs (Barnes, 1989; Buchanan & Heinrichs, 1989; Chouinard & Margolese, 2005; Guy, 1976; Sanders et al., 2005; Simpson & Angus, 1970). In clinical settings, these scales are often completed by physicians who are concerned with medication side effects or neurological involvement. With appropriate training, psychologists may reliably complete such scales to gain insight into medication side effects which cause personal distress, impaired functioning, and medication noncompliance, among other clinically relevant phenomena. Thus, there are many useful assessment procedures that clinicians may use to provide valuable insights into the symptoms and functioning of individuals with mental disorders.

## Acknowledgments

This work represents a collection of efforts from many people. The authors would like to extend special thanks to Jayson Wright, Anita Kwong, and Sarah Flood for their contributions to this chapter.

## References

- Addington, D., Addington, J., & Maticka-Tyndale, E. (1993). Assessing depression in schizophrenia: The Calgary depression scale. *British Journal of Psychiatry*, 163(Suppl. 22 (S22)), 39–44.
- Addington, D., Addington, J., & Maticka-Tyndale, E. (1994). Specificity of the Calgary depression scale for schizophrenics. *Schizophrenia Research*, 11(3), 239–244.
- Addington, D., Addington, J., & Schissel, B. (1990). A depression rating scale for schizophrenics. *Schizophrenia Research*, 3, 247–251.
- Ahmed, A. O., Kirkpatrick, B., Galderisi, S., Mucci, A., Maj, M., Rossi, A., ... Schneider, K. (in press). Cross cultural validation of the five-factor structure of negative symptoms in schizophrenia. *Schizophrenia Bulletin*. Available from <https://doi.org/10.1093/schbul/sby050>.
- Alegria, M., Jackson, J. S., Kessler, R. C., & Takeuchi, D. (retrieved 6/1/18). *Collaborative Psychiatric Epidemiology Surveys (CPES), 2001–2003 [United States]*. Ann Arbor, MI, Inter-University Consortium for Political and Social Research [distributor], 2016-03-23. Available from <https://doi.org/10.3886/ICPSR20240.v8>.
- Alexopoulos, G. S., Abrams, R. C., Young, R. C., & Shamoian, C. A. (1988). Cornell scale for depression in dementia. *Biological Psychiatry*, 23, 271–284.

- Alloy, L. B., & Abramson, L. Y. (2010). The role of the behavioral approach system (BAS) in bipolar spectrum disorders. *Current Directions in Psychological Science*, 19(3), 189–194.
- Alberman, H., Snider, E. C., Cacciola, J. S., Brown, L. S., Jr., Zaballero, A., & Sidiqui, N. (1996). Evidence of response set effects in structured research interviews. *Journal of Nervous and Mental Disease*, 184, 403–410.
- American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Arlington, VA: American Psychiatric Publishing.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- Andreasen, N. A., Arndt, S., Alliger, R., Miller, D., & Flaum, M. (1995). Symptoms of schizophrenia: Methods, meanings, and mechanisms. *Archives of General Psychiatry*, 52, 341–351.
- Andreasen, N. C. (1983). *Scale for the assessment of negative symptoms (SANS)*. Iowa City, IA: University of Iowa.
- Andreasen, N. C. (1984). *Scale for the assessment of positive symptoms (SANS)*. Iowa City, IA: University of Iowa.
- Andreasen, N. C., & Olsen, S. (1982). Negative vs. positive schizophrenia: Definition and validation. *Archives of General Psychiatry*, 39, 789–794.
- Axelsson, E., Andersson, E., Ljótsson, B., Wallhed Finn, D., & Hedman, E. (2016). The health preoccupation diagnostic interview: Inter-rater reliability of a structured interview for diagnostic assessment of DSM-5 somatic symptom disorder and illness anxiety disorder. *Cognitive Behaviour Therapy*, 45(4), 259–269.
- Ballard, E. D., Yarrington, J. S., Farmer, C. A., Lener, M. S., Kadriu, B., Lally, N., ... Zarate, C. A. (2018). Parsing the heterogeneity of depression: An exploratory factor analysis across commonly used depression rating scales. *Journal of Affective Disorders*, 231, 51–57. Available from <https://doi.org/10.1016/j.jad.2018.01.027>.
- Barnes, T. R. (1989). A rating scale for drug-induced akathisia. *The British Journal of Psychiatry*, 154(5), 672–676.
- Bech, P., Bolwig, T. G., Kramp, P., & Rafaelsen, O. J. (1979). The Bech–Rafaelsen mania scale and the Hamilton depression scale. *Acta Psychiatrica Scandinavica*, 59(4), 420–430.
- Bech, P. (2002). The Bech–Rafaelsen mania scale in clinical trials of therapies for bipolar disorder: A 20-year review of its use as an outcome measure. *CNS Drugs*, 16(1), 47–63.
- Bech, P., Bastrup, P. C., de Bleeker, E., & Ropert, R. (2001). Dimensionality, responsiveness and standardisation of the Bech–Rafaelsen mania scale in the ultra-short therapy with antipsychotics in patients with severe manic episodes. *Acta Psychiatrica Scandinavica*, 104(1), 25–30.
- 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.
- Beigel, A., & Murphy, D. L. (1971). Assessing clinical characteristics of the manic state. *American Journal of Psychiatry*, 128(6), 688–694.
- Bigelow, L., & Murphy, D.L. (1978). *Guidelines and anchor points for modified BPRS* (unpublished manuscript). NIMH Intramural Research Program. Washington, DC: Saint Elizabeth's Hospital.
- Bilder, R. M., Mukherjee, S., & Rieder, R. O. (1985). Symptomatic and neuropsychological components of defect state. *Schizophrenia Bulletin*, 11(3), 409–419.
- Birchwood, M., Smith, J., Cockrane, R., Wetton, S., & Copestake, S. (1990). The social functioning scale: The development and validation of a new scale of social adjustment for use in family intervention programmes with schizophrenic patients. *British Journal of Psychiatry*, 157, 853–859.

- Black, D. W., Arndt, S., Hale, N., & Rogerson, R. (2004). Use of the mini international neuro-psychiatric interview (MINI) as a screening tool in prisons: Results of a preliminary study. *Journal of the American Academy of Psychiatry and the Law Online*, 32(2), 158–162.
- Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Gusman, F. D., Charney, D. S., & Keane, T. M. (1995). The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress*, 8, 75–90.
- Blanchard, E. P., Jones-Alexander, J., Buckley, T. C., & Forneris, C. A. (1996). Psychometric properties of the PTSD Checklist. *Behaviour Research and Therapy*, 34, 669–673.
- Blanchard, J. J., & Cohen, A. S. (2006). The structure of negative symptoms within schizophrenia: Implications for assessment. *Schizophrenia Bulletin*, 32(2), 238–245.
- Brown, C., Schulberg, H. C., & Madonia, M. J. (1995). Assessing depression in primary care practice with the Beck depression inventory and the Hamilton rating scale for depression. *Psychological Assessment*, 7, 59–65.
- Brugha, T. S., Bebbington, P. E., Jenkins, R., Meltzer, H., Taub, N. A., Janas, M., & Vernon, J. (1999). Cross validation of a general population survey diagnostic interview: A comparison of CIS-R with SCAN ICD-10 diagnostic categories. *Psychological Medicine*, 29 (5), 1029–1042. Available from <https://doi.org/10.1017/S0033291799008892>.
- Buchanan, R. W., & Heinrichs, D. W. (1989). The neurological evaluation scale (NES): A structured instrument for the assessment of neurological signs in schizophrenia. *Psychiatry Research*, 27, 335–350.
- Carpenter, W. T., Jr., Heinrichs, D. W., & Wagman, A. M. (1988). Deficit and nondeficit forms of schizophrenia: The concept. *American Journal of Psychiatry*, 145(5), 578–583.
- Carpenter, W. T., Conley, R. R., Buchanan, R. W., Breier, A., Tamminga, C. A., et al. (1995). Patient response and resource management: Another view of clozapine treatment of schizophrenia. *American Journal of Psychiatry*, 152(6), 827–832.
- Cassidy, F., Forest, F., Murry, E., & Carroll, B. J. (1998). A factor analysis of the signs and symptoms of mania. *Archives of General Psychiatry*, 55, 27–32. Available from <https://doi.org/10.1001/archpsyc.55.1.27>.
- Cassidy, F., Murry, E., Forest, K., & Carroll, B. J. (1998). Signs and symptoms of mania in pure and mixed episodes. *Journal of Affective Disorders*, 50(2–3), 187–201. Available from [https://doi.org/10.1016/S0165-0327\(98\)00016-0](https://doi.org/10.1016/S0165-0327(98)00016-0).
- Chen, P. J. (1990). The efficacy and blood concentration monitoring of carbamazepine on mania (in Chinese). *Zhonghua Shen Jing Jing Shen Ke Za Zhi*, 23(5), 261–265, 318.
- Chouinard, G., & Margolese, H. C. (2005). Manual for the extrapyramidal symptom rating scale (ESRS). *Schizophrenia Research*, 76, 247–265.
- Colom, F., Vieta, E., Martinez-Aran, A., Garcia-Garcia, M., Reinares, M., Torrent, C., ... Salamero, M. (2002). Spanish version of a scale for the assessment of mania: Validity and reliability of the young mania rating scale. *Medicina Clinica*, 119(10), 366–371. Available from [https://doi.org/10.1016/S0025-7753\(02\)73419-2](https://doi.org/10.1016/S0025-7753(02)73419-2).
- Columbia DISC Development Group. (1999). *National Institute of Mental Health Diagnostic Interview Schedule for Children (NIMH-DISC)* (unpublished report). Columbia University/New York State Psychiatric Institute.
- Cooper, L., Peters, L., & Andrews, G. (1998). Validity of the composite international diagnostic interview (CIDI) psychosis module in a psychiatric setting. *Journal of Psychiatric Research*, 32(6), 361–368.
- Cottler, L. B., Grant, B. F., Blaine, J., Mavreas, V., Pull, C., Hasin, D., Compton, W. M., Rubio-Stipe, M., & Mager, D. (1997). Concordance of DSM-IV alcohol and drug use disorder criteria and diagnoses as measured by AUDADIS-ADR, CIDI and SCAN. *Drug and Alcohol Dependence*, 47, 195–205.

- Cottler, L. B., Compton, W. M., Ridenour, A., Abdahlla, A. B., & Gallagher, T. (1998). Reliability of self-reported antisocial personality disorder symptoms among substance abusers. *Drug and Alcohol Dependence*, 49, 189–199.
- Davidson, J. R. T., Miner, C. M., De Veaugh Geiss, J., Tupler, L. A., Colket, J. T., & Potts, N. L. S. (1997). The brief social phobia scale: A psychometric evaluation. *Psychological Medicine*, 27, 161–166.
- Davidson, J., Turnbull, C. D., Strickland, R., Miller, R., & Graves, K. (1986). The Montgomery–Asberg depression scale: Reliability and validity. *Acta Psychiatrica Scandinavica*, 73, 544–548.
- Davis, L., Uezato, A., Newell, J. M., & Frazier, E. (2008). Major depression and comorbid substance use disorders. *Current Opinions in Psychiatry*, 21(1), 14–18. Available from <https://doi.org/10.1097/YCO.0b013e3282f32408>.
- Diamond, R., & Becker, M. (1999). The Wisconsin quality of life index: A multidimensional model for measuring quality of life. *Journal of Clinical Psychiatry*, 60(Suppl. 3), 29–31.
- Dilsaver, S. C., Chen, R., Shoaib, A. M., & Swann, A. C. (1999). Phenomenology of mania: Evidence for distinct depressed, dysphoric, and euphoric presentations. *American Journal of Psychiatry*, 156(3), 426–430.
- Dingemans, P. M. A. J., Linszen, D. H., Lenior, M. E., & Smeets, R. M. W. (1995). Component structure of the expanded brief psychiatric rating scale (BPRS-E). *Psychopharmacology*, 122(3), 263–267.
- Easton, C., Meza, E., Mager, D., Ulug, B., Kilic, C., Gogus, A., & Babor, T. F. (1997). Test–retest reliability of the alcohol and drug use disorder sections of the schedules for clinical assessment in neuro-psychiatry (SCAN). *Drug and Alcohol Dependence*, 47(3), 187–194. Available from [https://doi.org/10.1016/S0376-8716\(97\)00089-6](https://doi.org/10.1016/S0376-8716(97)00089-6).
- Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The schedule for affective disorders and schizophrenia. *Archives of general psychiatry*, 35(7), 837–844.
- Fairburn, C. G., & Cooper, Z. (1993). The eating disorder examination. In C. G. Fairburn, & G. T. Wilson (Eds.), *Binge eating: Nature, assessment, and treatment* (12th ed., pp. 317–360). New York, NY: Guilford Press.
- Farmer, A., Chubb, H., Jones, I., Hillier, J., Smith, A., & Borysiewicz, L. (1996). Screening for psychiatric morbidity in subjects presenting with chronic fatigue syndrome. *British Journal of Psychiatry*, 168(3), 354–358. Available from <https://doi.org/10.1192/bjp.168.3.354>.
- Farmer, A., Cosyns, P., Leboyer, M., Maier, W., Mors, O., Sargeant, M., ... McGuffin, P. (1993). A SCAN–SADS comparison study of psychotic subjects and their 1st-degree relatives. *European Archives of Psychiatry and Clinical Neuroscience*, 242(6), 352–356. Available from <https://doi.org/10.1007/Bf02190248>.
- Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry*, 26, 57–63.
- Fields, J. H., Grochowski, S., Lindenmayer, J. P., Kay, S. R., Grosz, D., Hyman, R. B., & Alexander, G. (1994). Assessing positive and negative symptoms in children and adolescents. *American Journal of Psychiatry*, 151(2), 249–253.
- First, M. B., Skodol, A. E., Bender, D. S., & Oldham, J. M. (2018). *User's guide for SCID-5-AMPD*. Arlington, VA: American Psychiatric Publishing.
- First, M. B., Williams, J. B. W., Benjamin, L. S., & Spitzer, R. L. (2016). *Structured clinical interview for DSM-5® personality disorders (SCID-5-PD)*. Arlington, VA: American Psychiatric Publishing.

- First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2015a). *Structured clinical interview for DSM-5—Research version (SCID-5 for DSM-5, research version; SCID-5-RV)*. Arlington, VA: American Psychiatric Association.
- First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2015b). *Structured clinical interview for DSM-5 disorders, clinician version (SCID-5-CV)*. Arlington, VA: American Psychiatric Association.
- First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2015c). Structured clinical interview for DSM-5 disorders, clinical trials version (SCID-5-CT). Arlington, VA: American Psychiatric Association.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method of grading cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189–198.
- Fristad, M. A., Weller, R. A., & Weller, E. B. (1995). The mania rating scale (MRS): Further reliability and validity studies with children. *Annals of Clinical Psychiatry*, 7, 127–132.
- Galderisi, S., Mucci, A., Bitter, I., Libiger, J., Bucci, P., Fleischhacker, W. W., ... Eufest Study Group. (2013). Persistent negative symptoms in first episode patients with schizophrenia: Results from the European first episode schizophrenia trial. *European Neuropsychopharmacology*, 23(3), 196–204.
- Gjerris, A., Bech, P., Broen-Christensen, C., Geisler, A., Klysner, R., & Rafaelsen, O. J. (1980). Haloperidol plasma levels in relation to antimanic effect. In E. Usdin, S. G. Dahl, L. F. Gram, & O. Lingjaerde (Eds.), *Clinical pharmacology in psychiatry* (pp. 227–232). London: McMillan.
- Gozdzik-Zelazny, A., Borecki, L., & Pokorski, M. (2011). Depressive symptoms in schizophrenic patients. *European Journal of Medical Research*, 16, 549–5210. Available from <https://doi.org/10.1186/2047-783X-16-12-549>.
- Gunter, T. D., Arndt, S., Wenman, G., Allen, J., Loveless, P., Sieleni, B., & Black, D. W. (2008). Frequency of mental and addictive disorders among 320 men and women entering the Iowa prison system: Use of the MINI-Plus. *Journal of the American Academy of Psychiatry and the Law Online*, 36(1), 27–34.
- Guy, W. A. (1976). *Abnormal involuntary movement scale (AIMS). ECDEU assessment manual for psychopharmacology* (pp. 534–537). Washington, DC: Department of Health Education and Welfare.
- Hafkenscheid, A. (1993). Reliability of a standardized and expanded brief psychiatric rating scale: A replication study. *Acta Psychiatrica Scandanavica*, 88, 305–310.
- Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology*, 32, 50–55.
- Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology Neurosurgery and Psychiatry*, 23, 56–62.
- Hamilton, M. (2000). *Hamilton rating scale for depression (Ham-D). Handbook of psychiatric measures* (pp. 526–528). Washington, DC: American Psychiatric Association.
- Hanwella, R., & de Silva, V. A. (2011). Signs and symptoms of acute mania: A factor analysis. *BMC Psychiatry*, 11, 137. Available from <https://doi.org/10.1186/1471-244X-11-137>.
- Hapke, U., Rumpf, H. J., & John, U. (1998). Differences between hospital patients with alcohol problems referred for counselling by physicians' routine clinical practice versus screening questionnaires. *Addiction*, 93(12), 1777–1785. Available from <https://doi.org/10.1046/j.1360-0443.1998.931217774>.
- Haro, J. M., Arbabzadeh-Bouchez, S., Brugha, T. S., De Girolamo, G., Guyer, M. E., Jin, R., Lepine, J. P., et al. (2006). Concordance of the composite international diagnostic interview version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World

- Mental Health Surveys. *International Journal of Methods in Psychiatric Research*, 15(4), 167–180.
- Harvey, P. D., Endicott, J. M., & Loebel, A. D. (2008). The factor structure of clinical symptoms in mixed and manic episodes prior to and after antipsychotic treatment. *Bipolar Disorders*, 10(8), 900–906. Available from <https://doi.org/10.1111/j.1399-5618.2008.00634>.
- Hegeman, J. M., Wardenaar, K. J., Comijs, H. C., de Waal, M. W. M., Kok, R. M., & van der Mast, R. C. (2012). The subscale structure of the inventory of depressive symptomatology self report (IDS-SR) in older persons. *Journal of Psychiatric Research*, 46(10), 1383–1388. Available from <https://doi.org/10.1016/j.jpsychires.2012.07.008>.
- Heinrichs, D. W., Hanlon, T. E., & Carpenter, W. T., Jr. (1984). The quality of life scale: An instrument for rating the schizophrenic deficit syndrome. *Schizophrenia Bulletin*, 10(3), 388–398.
- Helzer, J. E., Robins, L. N., McEnvoy, L. T., Spitznagel, L. M., Stoltzman, R. K., Farmer, A., & Brockington, I. F. (1985). A comparison of clinical and diagnostic interview schedule diagnoses. *Archives of General Psychiatry*, 42, 657–666.
- Hesselbrock, V., Stabenau, J., Hesselbrock, M., Mirkin, P., & Meyer, R. (1982). A comparison of two interview schedules: The S for affective disorders and schizophrenia-lifetime and national institute of mental health diagnostic interview schedule. *Archives of General Psychiatry*, 39, 674–677.
- Huxley, P., Korer, J., & Tolley, S. (1987). The psychiatric “caseness” of clients referred to an urban social services department. *British Journal of Sociology*, 17, 507–520.
- Inch, R., Crossley, M., Keegan, D., & Thorarinson, D. (1997). Use of the brief psychiatric rating scale to measure success in a psychosocial day program. *Psychiatric Services*, 48(9), 1195–1197.
- Jaeger, J., Berns, S. M., & Czobor, P. (2003). The multidimensional scale of independent functioning: A new instrument for measuring functional disability in psychiatric populations. *Schizophrenia Bulletin*, 29(1), 153–168.
- Janca, A., Ustun, T. B., & Sartorius, N. (1994). New version of World Health Organization instruments for assessment of mental disorders. *Acta Psychiatrica Scandinavica*, 90, 440–443.
- Johnson, S. L., Cuellar, A. K., Ruggero, C., Winett-Perlman, C., Goodnick, P., White, R., et al. (2008). Life events as predictors of mania and depression in bipolar I disorder. *Journal of Abnormal Psychology*, 117, 268–277.
- Johnson, S. L., Edge, M. D., Holmes, M. K., & Carver, C. S. (2012). The behavioral activation system and mania. *Annual Review of Clinical Psychology*, 8, 243–267.
- Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, 13(2), 261–276.
- Kelley, M. E., Gilbertson, M. W., Mouton, A., & van Kammen, D. P. (1992). Deterioration in premorbid functioning in schizophrenia: A developmental model of negative symptoms in drug-free patients. *American Journal of Psychiatry*, 149, 1543–1548.
- Kessler, R. C., & Üstün, T. B. (2004). The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*, 13(2), 93–121.
- Kirkpatrick, B., Buchanan, R. W., McKenney, P. D., Alphs, L. D., & Carpenter, W. T. (1989). The schedule for the deficit syndrome: An instrument for research in schizophrenia. *Psychiatry Research*, 30, 119–123.

- Kirkpatrick, B., Fenton, W. S., Carpenter, W. T., & Marder, S. R. (2006). The NIMH-MATRICS consensus statement on negative symptoms. *Schizophrenia Bulletin*, 32, 214–219.
- Kirkpatrick, B., Strauss, G. P., Nguyen, L., Fischer, B. A., Daniel, D. G., Cienfuegos, A., ... Marder, S. R. (2011). The brief negative symptom scale: Psychometric properties. *Schizophrenia Bulletin*, 37(2), 300–305. Available from <https://doi.org/10.1093/schbul/sbq059>.
- Kovess, V., & Fournier, L. (1990). The DISSA: An abridged self-administered version of the DIS: Approach by episode. *Social Psychiatry and Psychiatric Epidemiology*, 25(4), 179–186.
- Kraepelin, E. (1921). *Manic-depressive insanity and paranoia* (R. M. Barclay, Trans.). In G. M. Robertson (Ed.). Edinburgh: E & S Livingstone.
- Lachar, D., Bailley, S. E., Rhoades, H. M., Espadas, A., Aponte, M., Cowan, K. A., ... Wassef, A. (2001). New subscales for an anchored version of the brief psychiatric rating scale: Construction, reliability, and validity in acute psychiatric admissions. *Psychological Assessment*, 13, 384–395.
- Lako, I. M., Bruggeman, R., Knegtering, H., Wiersma, D., Schoevers, R. A., Slooff, C. J., ... Taxis, K. (2012). A systematic review of instruments to measure depressive symptoms in patients with schizophrenia. *Journal of Affective Disorders*, 140(1), 38–47.
- Lesage, A. D., Cyr, M., & Toupin, J. (1991). Reliable use of the present state examination by psychiatric nurses for clinical studies of psychotic and nonpsychotic patients. *Acta Psychiatrica Scandinavica*, 83, 121–124.
- Licht, R. W., & Jensen, J. (1997). Validation of the Bech–Rafaelsen mania scale using latent structure analysis. *Acta Psychiatrica Scandinavica*, 96, 367–372.
- Liebowitz, M. R. (1987). Social phobia. *Modern Problems in Pharmacopsychiatry*, 22, 141–173.
- Lindenmayer, J. P., Harvey, P. D., Khan, A., & Kirkpatrick, B. (2007). Schizophrenia: Measurements of psychopathology. *Psychiatric Clinics of North America*, 30(3), 339–363.
- 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 & Psychotherapy*, 8(1), 75–79.
- Long, J. D., & Brekke, J. S. (1999). Longitudinal factor structure of the brief psychiatric rating scale in schizophrenia. *Psychological Assessment*, 11, 498–506.
- Loranger, A. W., Janca, A., & Sartorius, N. (1997). *Assessment and diagnosis of personality disorders. The CD-10 international personality disorder examination (IPDE)*. Cambridge: Cambridge University Press.
- Loranger, A. W., Sartorius, N., Andreoli, A., Berger, P., Buchheim, P., Channabasavanna, S. M., ... Regier, D. A. (1994). The international personality disorder examination: The World Health Organization/alcohol, drug abuse, and mental health administration international pilot study of personality disorders. *Archives of General Psychiatry*, 51(3), 215.
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... Rutter, M. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3), 205–223.
- Lukoff, D., Nuechterlein, K. H., & Ventura, J. (1986). Appendix A: Manual for expanded brief psychiatric rating scale (BPRS). *Schizophrenia Bulletin*, 12, 594–602.
- Maffei, C., Fossati, A., Agostoni, I., Barraco, A., Bagnato, M., Deborah, D., & Petrachi, M. (1997). Interrater reliability and internal consistency of the structured clinical interview

- for DSM-IV Axis II personality disorders (SCID-II), version 2.0. *Journal of Personality Disorders*, 11(3), 279–284.
- Maggini, C., & Raballo, A. (2006). Exploring depression in schizophrenia. *European Psychiatry*, 21, 227–232.
- Malkoff-Schwartz, S., Frank, E., Anderson, B., Sherrill, J. T., Siegel, L., Patterson, D., et al. (1998). Stressful life events and social rhythm disruption in the onset of manic and depressive bipolar episodes. *Archives of General Psychiatry*, 55, 702–707.
- Marcos, T., & Salamero, M. (1990). Factor study of the Hamilton rating scale for depression and the Bech Melancholia scale. *Acta Psychiatrica Scandinavica*, 82, 178–181.
- Marneros, A., & Angst, J. (2002). *Bipolar disorders: Roots and evolution. Bipolar disorders: 100 years after manic depressive insanity* (pp. 1–36). Springer, pp.
- Mausbach, B. T., Harvey, P. D., Goldman, S. R., Jeste, D. V., & Patterson, T. L. (2007). Development of a brief scale of everyday functioning in persons with serious mental illness. *Schizophrenia Bulletin*, 33(6), 1364–1372.
- McLellan, A. T., Kushner, H., & Metzger, D. (1992). The fifth edition of the addiction severity index. *Journal of Substance Abuse Treatment*, 9(3), 199–213.
- Mignoli, G., Facciniani, C., Turit, L., Gavioli, I., & Micciolo, R. (1988). Interrater reliability of the PSE-9 (full version): An Italian study. *Social Psychiatry and Psychiatric Epidemiology*, 23, 30–35.
- Moberg, P. J., Lazarus, L. W., Mesholam, R. I., Bilker, W., Chuy, I. L., Neyman, I., & Markvart, V. (2001). Comparison of the standard and structured interview guide for the Hamilton Depression Rating Scale in depressed geriatric inpatients. *American Journal of Geriatric Psychiatry*, 9, 35–40.
- Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, 134, 382–389.
- Moran, K. K., & Tan, S. (1998). Comparison of the brief psychiatric rating scale and the brief symptom inventory. *Journal of Clinical Psychology*, 54(7), 885–894.
- Mueser, K. T., Curran, P. J., & McHugo, G. J. (1997). Factor structure of the brief psychiatric rating scale in schizophrenia. *Psychological Assessment*, 9(3), 196–204.
- National Institute of Mental Health. (1991). NIMH diagnostic interview for children, version 2.3. Rockville, MD: Author.
- Nicholson, I. R., Chapman, J. E., & Neufeld, R. W. J. (1995). Variability in BPRS definitions of positive and negative symptoms. *Schizophrenia Research*, 17, 177–185.
- Norman, R. M. G., Malla, A. K., Cortese, L., & Diaz, J. F. (1996). A study of the interrelationships between and comparative inter-rater reliability of SAPS, SANS and PANSS. *Schizophrenia Research*, 19, 73–85.
- North, C. S., Pollio, D. E., Thompson, S. J., Ricci, D. A., Smith, E. M., & Spitznagel, E. L. (1997). A comparison of clinical and structured interview diagnoses in a homeless mental health clinic. *Community Mental Health Journal*, 33(6), 531–543.
- Oldham, J. M., Skodol, A. E., Kellman, H. D., Hyler, S. E., Doidge, N., Rosnick, L., & Gallaher, P. E. (1995). Comorbidity of axis I and axis II disorders. *American Journal of Psychiatry*, 152(4), 571–578.
- Overall, J. E., & Gorham, D. R. (1962). The brief psychiatric rating scale. *Psychological Reports*, 10, 799–812.
- Overall, J. E., & Pfefferbaum, B. (1982). The brief psychiatric rating scale for children. *Psychopharmacology Bulletin*, 18, 10–16.
- Overall, J. E., Hollister, L. E., & Pichot, P. (1967). Major psychiatric disorders: A four-dimensional model. *Archives of General Psychiatry*, 16, 146–151.

- Patterson, T. L., Goldman, S., McKibbin, C. L., Hughs, T., & Jeste, D. V. (2001). UCSD performance-based skills assessment: Development of a new measure of everyday functioning for severely mentally ill adults. *Schizophrenia Bulletin*, 27(2), 235–245.
- Peralta, V., & Cuesta, M. J. (1999). Dimensional structure of psychotic symptoms: An item-level analysis of the SAPS and SANS symptoms in psychotic disorders. *Schizophrenia Research*, 38, 13–26.
- Peralta, V., & Cuesta, M. J. (2001). How many and which are the psychopathological dimensions in schizophrenia? Issues influencing their ascertainment. *Schizophrenia Research*, 49(3), 269–285.
- Peveler, R. C., & Faiburn, C. G. (1990). Measurement of neurotic symptoms by self-report questionnaire: Validity of the SCL-90R. *Psychological Medicine*, 20, 873–879.
- Presnell, K., & Stice, E. (2003). An experimental test of the effect of weight-loss dieting on bulimic pathology: Tipping the scales in a different direction. *Journal of Abnormal Psychology*, 112, 166–170.
- Quilty, L. C., Zhang, K. A., & Bagby, R. M. (2010). The latent symptom structure of the Beck Depression Inventory-II in outpatients with major depression. *Psychological Assessment*, 22(3), 603–608.
- Rabany, L., Weiser, M., & Levkovitz, Y. (2013). Guilt and depression: Two different factors in individuals with negative symptoms of schizophrenia. *European Psychiatry*, 28(6), 327–331.
- Regier, D. A., Myers, J. K., Kramer, M., Robins, L. N., Blazer, D. G., Hough, R. L., ... Locke, B. Z. (1984). The NIMH epidemiological catchment area program: Historical context, major objectives, and study population characteristics. *Archives of General Psychiatry*, 41(10), 934–941.
- Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, 170(1), 59–70.
- Rekhi, G., Ng, W. Y., & Lee, J. (2018). Clinical utility of the Calgary depression scale for schizophrenia in individuals at ultra-high risk of psychosis. *Schizophrenia Research*, 193, 423–427.
- Robertson, G. M. (1890). Does mania include two distinct varieties of insanity and should it be subdivided? *Journal of Mental Science*, 36, 338–347.
- Robins, L. N., Cottler, L. B., & Keating, S. (1991). *NIMH diagnostic interview schedule, version III-revised (DIS-III-R): Question by question specifications*. St. Louis, MO: Washington School of Medicine.
- Robins, L. N., Cottler, L. B., Bucholz, K., & Compton, W. (1995). *NIMH diagnostic interview schedule, version IV*. St. Louis, MO: Washington School of Medicine.
- Rodriguez-Jimenez, R., Bagney, A., Mezquita, L., Martinez-Grasa, I., & Sanchez-Morla, E. (2013). Cognition and the five-factor model of the positive and negative syndrome scale in schizophrenia. *Schizophrenia Research*, 143(1), 77–83.
- Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. New York, NY: Guilford Press.
- Roncone, R., Ventura, J., Impallomeni, M., Fallon, I. R. H., & Morosini, P. L. (1999). Reliability of an Italian standardized and expanded brief psychiatric rating scale (BPRS 4.0) in raters with high vs. low clinical experience. *Acta Psychiatrica Scandinavica*, 100, 229–236.

- Rossi, A., Daneluzzo, E., Arduini, L., Di Domenico, M., Pollice, R., & Petruzz, C. (2001). A factor analysis of signs and symptoms of the manic episode with Bech–Rafaelsen mania and melancholia scales. *Journal of Affective Disorders*, 64, 267–270.
- Rush, A. J., First, M. B., & Blacker, D. (2008). *Handbook of Psychiatric Measures* (2nd ed.). Arlington, VA: American Psychiatric Publishing.
- Rush, A. J., Giles, D. E., Schlessier, M. A., Fulton, C. L., Weissenburger, J., & Burns, C. (1986). The inventory for depressive symptomatology (IDS): Preliminary findings. *Psychiatry Research*, 18, 65–87.
- Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The inventory of depressive symptomatology (IDS): Psychometric properties. *Psychological Medicine*, 26, 477–486.
- Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., ... Keller, M. B. (2003). The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDSSR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54, 573–583.
- Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism diagnostic interview—revised*. Los Angeles, CA: Western Psychological Services.
- Sanders, R. D., Allen, D. N., Forman, S. D., Tarpey, T., Keshavan, M. S., & Goldstein, G. (2005). Confirmatory factor analysis of the neurological evaluation scale in unmedicated schizophrenia. *Psychiatry Research*, 133, 65–71.
- Schennach, R., Obermeier, M., Seemüller, F., Jager, M., Schmauss, M., Laux, G., ... Gaebel, W. (2012). Evaluating depressive symptoms in schizophrenia: A psychometric comparison of the Calgary depression scale for schizophrenia and the Hamilton depression rating scale. *Psychopathology*, 45(5), 276–285.
- Schneider, L. C., & Struening, E. L. (1983). SLOF: A behavioral rating scale for assessing the mentally ill. *Social Work Research & Abstracts*, 19(3), 9–21.
- Scott, J., Murray, G., Henry, C., Morken, G., Scott, E., Angst, J., ... Hickie, I. B. (2017). Activation in bipolar disorders: A systematic review. *JAMA Psychiatry*, 74(2), 189–196.
- Shafer, A. B. (2005). Meta-analysis of the brief psychiatric rating scale factor structure. *Psychological Assessment*, 17(3), 324–335.
- Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, 62(1), 123–146.
- Shafer, A. B. (2013). Factor structure of the brief psychiatric rating scale for children (BPRS-C) among hospital patients and community clients. *Personality and Individual Differences*, 55(1), 41–46.
- Shankman, S. A., Funkhouser, C. J., Klein, D. N., Davila, J., Lerner, D., & Hee, D. (2018). Reliability and validity of severity dimensions of psychopathology assessed using the structured clinical interview for DSM-5 (SCID). *International Journal of Methods in Psychiatric Research*, 27(1), e1590.
- Shansis, F. M., Reche, M., & Capp, E. (2016). Evaluating response to mood stabilizers in patients with mixed depression: A study of agreement between three different mania rating scales and a depression rating scale. *Journal of Affective Disorders*, 197, 1–7.
- Sheehan, D. V., Lecrubier, Y., Harnett Sheehan, K., Janavs, J., Weiller, E., Keskiner, A., & Dunbar, G. C. (1997). The validity of the mini international neuropsychiatric interview (MINI) according to the SCID-P and its reliability. *European Psychiatry*, 12(5), 232–241.
- Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., & Dunbar, G. C. (1998). The mini-international neuropsychiatric interview (MINI): The

- development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, 59, 22–33.
- Sheehan, D., Janavs, J., Baker, R., Sheehan, K. H., Knapp, E., & Sheehan, M. (2015). *Mini international neuropsychiatric interview—version 7.0.0 DSM-5. 2014*.
- Simpson, G. M., & Angus, J. W. S. (1970). A rating scale for extrapyramidal side effects. *Acta Psychiatrica Scandinavica*, 212(Suppl.), 11–19.
- Singh, M. M., & Kay, S. R. (1987). Is the positive–negative distinction in schizophrenia valid? *British Journal of Psychiatry*, 150, 879–880.
- Sobell, M. B., Sobell, L. C., Klajner, F., Pavan, D., & Basian, E. (1986). The reliability of the timeline method of assessing normal drinker college students' recent drinking history: Utility of alcohol research. *Addictive Behaviors*, 11, 149–161.
- Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria. *Archives of General Psychiatry*, 35, 773–782.
- Strauss, G. P., Hong, L. E., Keller, W. R., Buchanan, R. W., & Gold, J. M. (2012). Factor structure of the brief negative symptom scale. *Schizophrenia Research*, 142, 96–98.
- Strauss, G. P., Nuñez, A., Ahmed, A. O., Barchard, K. A., Granholm, E., Kirkpatrick, B., & Allen, D. N. (2018). The latent structure of negative symptoms in schizophrenia. *JAMA Psychiatry*, 75, 1271–1279.
- Substance Abuse and Mental Health Services Administration. (2017). *Key substance use and mental health indicators in the United States: Results from the 2016 national survey on drug use and health*. Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, HHS Publication No. SMA 17-5044, NSDUH Series H-52.
- Sunderland, M., Carragher, N., Wong, N., & Andrews, G. (2013). Factor mixture analysis of DSM-IV symptoms of major depression in a treatment seeking clinical population. *Comprehensive Psychiatry*, 54, 474–483.
- Swann, A. C., Lafer, B., Perugi, G., Frye, M. A., Bauer, M., Bahk, M. W., & Suppes, T. (2013). Bipolar mixed states: An international society for bipolar disorders task force report of symptom structure, course of illness, and diagnosis. *American Journal of Psychiatry*, 170, 31–42.
- Takahashi, N., Tomita, K., Higuchi, T., & Inada, T. (2004). The inter-rater reliability of the Japanese version of the Montgomery–Asberg depression rating scale (MADRS) using a structured interview guide for MADRS (SIGMA). *Human Psychopharmacology*, 19, 187–192.
- Trivedi, M. H., Rush, A. J., Ibrahim, H. M., Carmody, T. J., Biggs, M. M., Suppes, T., et al. (2004). The inventory of depressive symptomatology, clinician rating (IDS-C) and self-report (IDS-SR), and the quick inventory of depressive symptomatology, clinician rating (QIDS-C) and self-report (QIDS-SR) in public sector patients with mood disorders: A psychometric evaluation. *Psychological Medicine*, 34, 73–82.
- Van der Does, A. J., Dingemans, P. M., Linszen, D. H., Nugter, M. A., & Scholte, W. F. (1995). Dimensions and subtypes of recent-onset schizophrenia: A longitudinal analysis. *Journal of Nervous & Mental Disease*, 183, 681–687.
- Ventura, J., Green, M. F., Shaner, A., & Liberman, R. P. (1993). Training and quality assurance with the brief psychiatric rating scale: 'The drift busters'. *International Journal of Methods in Psychiatric Research*, 3, 221–244.
- Ward, L. C. (2006). Comparison of factor structure models for the Beck depression inventory-II. *Psychological Assessment*, 18, 81–88.

- Wardenaar, K. J., van Veen, T., Giltay, E. J., den Hollander-Gijsman, M., Penninx, B. W., & Zitman, F. G. (2010). The structure and dimensionality of the inventory of depressive symptomatology self-report (IDS-SR) in patients with depressive disorders and healthy controls. *Journal of Affective Disorders*, 125, 146–154.
- Weertman, A., Arntz, A., Dreessen, L., Velzen, C. V., & Vertommen, S. (2003). Short-interval test-retest interrater reliability of the Dutch version of the structured clinical interview for DSM-IV personality disorders (SCID-II). *Journal of Personality Disorders*, 17(6), 562–567.
- Wells, K. B., Burnam, A., Leake, B., & Robins, L. N. (1988). Agreement between face to face and telephone administered versions of the depression section of the NIMH diagnostic interview schedule. *Journal of Psychiatric Research*, 22, 207–220.
- Williams, J. B. (1988). A structured interview guide for the Hamilton depression rating scale. *Archives of General Psychiatry*, 45, 742–747.
- Williams, J. B., Kobak, K. A., Bech, P., Engelhardt, N., Evans, K., Lipsitz, J., et al. (2008). The GRID-HAMD: Standardization of the Hamilton depression rating scale. *International Clinical Psychopharmacology*, 23, 120–129.
- Wilimink, F. W., & Snijders, T. A. B. (1989). Polytomous logistic regression analysis of the general health questionnaire and the present state examination. *Psychological Medicine*, 19, 755–764.
- Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R., ... Sartorius, N. (1990). SCAN: Schedules for clinical assessment in neuropsychiatry. *Archives of General Psychiatry*, 47(6), 589–593.
- Wing, J. K., Birely, J. L. T., Cooper, J. E., Graham, P., & Isaacs, A. (1967). Reliability of a procedure for measuring and classifying present psychiatric state. *British Journal of Psychiatry*, 113, 499–515.
- Wing, J. K., Sartorius, N., Ustun, T., Bedirhan, T., World Health Organization., & WHO SCAN Advisory Committee. (1998). In J. K. Wing, N. Sartorius, & T. B. Ustun (Eds.), *Diagnosis and clinical measurement in psychiatry: A reference manual for SCAN*. Cambridge: Cambridge University.
- Wittchen, H. U. (1994). Reliability and validity studies of the WHO composite international diagnostic interview (CIDI): A critical-review. *Journal of Psychiatric Research*, 28(1), 57–84.
- World Health Organization. (1994). *Schedules for clinical assessment of neuropsychiatry (SCAN)*. Geneva: Author.
- World Health Organization. (2004). *The World Health Organization World Mental Health Composite International Diagnostic Interview (WHO WMH-CIDI)*. Geneva: Author.
- Yesavage, J. A. (1988). Geriatric depression scale. *Psychopharmacology Bulletin*, 24, 709–711.
- Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: Reliability, validity and sensitivity. *British Journal of Psychiatry*, 133(11), 429–435.
- Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (2000). *Young mania rating scale. Handbook of psychiatric measures* (pp. 540–542). Washington, DC: American Psychiatric Association.
- Young, R. C., Peasley-Miklus, C., & Shulberg, H. C. (2007). Mood ratings scales and the psychopathology of mania in old age: Selected applications and findings. In M. Sajatovic, & F. Blow (Eds.), *Bipolar disorders in later life*. Baltimore, MD: Johns Hopkins Press.

- Youngstrom, E. A., Danielson, C. K., Findling, R. L., Gracious, B. L., & Calabrese, J. R. (2002). Factor structure of the young mania rating scale for use with youths ages 5 to 17 years. *Journal of Clinical Child and Adolescent Psychology*, 31(4), 567–572.
- Zanello, A., Berthoud, L., Ventura, J., & Merlo, M. C. G. (2013). The brief psychiatric rating scale (version 4.0) factorial structure and its sensitivity in the treatment of outpatients with unipolar depression. *Psychiatry Research*, 210(2), 626–633.

## Further reading

- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Arlington, VA: American Psychiatric Publishing.
- Janca, A., & Helzer, J. (1990). DSM-III-R criteria checklist. *Diagnostic Interview Schedule Newsletter*, 7, 17.
- Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, I. 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.
- Zink, D., Lee, B., & Allen, D. N. (2015). *Structured and semi-structured clinical interviews available for use among African American clients: Cultural considerations in the diagnostic interview process. Guide to Psychological Assessment with African-Americans* (pp. 19–42). Springer.

# **Part VII**

# **Personality Assessment**

# Overview of multidimensional inventories of psychopathology with a focus on the MMPI-2

13

Carolyn L. Williams<sup>1</sup>, James N. Butcher<sup>1</sup> and Jacob A. Paulsen<sup>2</sup>

<sup>1</sup>Emeritus Professor, University of Minnesota, Minneapolis, MN, United States,

<sup>2</sup>San Diego City College, San Diego, CA, United States

## Precursors to self-report inventories of psychopathology

Throughout history, various methods have been used to classify or predict human behavior. Behavioral observations were commonly used in ancient civilizations and lie at the core of most informal strategies for understanding people. [Hathaway \(1965\)](#) pointed to the Old Testament for one of the earliest descriptions of the use of behavioral observations for assessing personality characteristics related to fitness for duty. According to Judges 7:4–8, God directed Gideon to observe his men before battle drinking water from a stream. Gideon was advised to select for his army the 300 men, out of a possible 10,000, who remained aware of their surroundings while drinking.

It was not until the 19th and early 20th centuries that more systematic procedures emerged to assess personality. These developments coincided with the beginnings of psychology as a profession. [Galton \(1884\)](#) wrote of assessing individual character through behavior observations. He suggested that questionnaires could be developed for measuring mental traits, although he never developed one. Later, [Cattell \(1890\)](#) coined the term “mental tests,” providing a scientific basis for the study of human behavior. In 1928 Hartshorne and May described a series of behavioral tests to measure “honesty.” The low correlations between the various single-situation honesty measures from their tests point out a basic principle in personality assessment: single items, whether they be behavioral observations or self-report items, tend to be unreliable predictors of individual characteristics or future behavior. Thus, *aggregation of information*, such as by grouping several observations across numerous situations, grouping information obtained by several different assessment methods, or simply adding individual items together to form scales, is basic to the development of reliable, valid, and predictive personality measures (e.g., [Anastasi, 1982](#); [Kruyken, Emons, & Sijtsma, 2012](#); [Kruyken, Emons, & Sijtsma, 2013](#); [Nunnally, 1978](#)).

The first formal use of a personality measure was by Heymans and Wiersma (1906). They developed a 90-item structured rating scale that was filled out by 3000 physicians about the patients they were treating. Contemporary self-report inventories soon overshadowed rating scales completed by individuals other than the subject being described. These self-report inventories contained aggregates of “conventional culturally crystallized questions to which the subject must respond in one of a few fixed ways” (Meehl, 1945, p. 296). Woodworth (1919, 1920) developed the first measure of adjustment that used such a series of written questions filled out by the individual.

The American Psychological Association (APA) commissioned Woodworth to develop a questionnaire to predict soldiers most likely to develop “shellshock” on World War I battlefields. Like Gideon’s behavioral observation technique described in the Old Testament, the first self-report instrument had its origins for military use. Woodworth’s Personal Data Sheet (PDS) included 116 face valid yes/no items including:

*“Do you feel well rested in the morning?”*

*“Is it easy to get you angry?”*

*“Have you ever fainted away?”*

*“Have you ever seen a vision?”*

The person’s score on the PDS was the total number of symptoms he or she endorsed. Much of the PDS item content can be found on self-report inventories in use today. There are a few items on the PDS illustrating how changes in language usage and the understanding of human behavior require item level modifications in self-report inventories (e.g., “Have you hurt yourself by masturbation [self-abuse]?” or “Did you ever have St. Vitus’ Dance?”) (Williams & Butcher, 2011). Woodworth (1919, 1920) compared samples of draftees and returning soldiers with “shellshock” with a sample of college students. However, the Personal Data Sheet was not finished in time to be used during World War I which ended in 1918.

Several types of self-report inventories evolved from this early work. These include multidimensional tests of normal personality like the California Psychological Inventory (CPI; Gough, 1956); the NEO Personality Inventory (NEO-PI; Costa & McCrae, 1985, 1992); and the Sixteen Personality Factors Questionnaire (16PF; Cattell, Cattell, & Cattell, 1993). Because they do not focus on psychopathology, they are not discussed in this chapter. In addition, there are a number of single dimension brief measures of psychopathology like the Beck Depression Inventory (BDI-II; Beck, Steer, & Brown, 1996) and the Beck Anxiety Inventory (BAI; Beck & Steer, 1990). The focus of this chapter is on the more comprehensive measures of psychopathology in current use: the MMPI-2, MMPI-2-RF, PAI, and MCMI. The emphasis will be on the most widely used instrument in both clinical practice and research, the MMPI-2.

## Development and use of multidimensional inventories of psychopathology

The most frequently used methods of scale construction for the development of self-report inventories of psychopathology are described in [Table 13.1](#). While developed in the 20th century, all continue to be used in the 21st century. During the 1930s and 1940s Hathaway and McKinley decided to follow the strategy of empirical scale construction based upon the work of [Paterson, Schneidler, and Williamson \(1938\)](#). In this method, often referred to as *criterion keying*, items were included on a scale if they empirically separated a clinical diagnostic group (e.g., depressed patients) from a sample of individuals without mental health problems (i.e., the “Minnesota Normal” sample). The resulting scales were the MMPI Clinical Scales. They are at the core of the most widely used multidimensional self-report inventories, the MMPI, the MMPI-2, and MMPI-A. Although [Hathaway and](#)

**Table 13.1** Scale construction methods for self-report inventories of psychopathology

*Rational or content-based item selection methods:* Early personality inventories relied on psychologists’ judgment, based on their understanding of psychopathology, for item selection. The rational approach assumes a direct correspondence between the item content and the personality attribute evaluated. Items typically have “face validity” (i.e., are easily understood to be related to the construct of interest). Rationally derived inventories can either be uni-dimensional or multidimensional. Woodworth’s Personal Data Sheet was uni-dimensional, measuring the broad construct “adjustment.” In multidimensional inventories items are combined into scales based on the developer’s insight into the item–construct relationships. They assess several personality traits or behavior patterns.

*Empirically derived methods.* In contrast to the selection of items based on the test developer’s theory, [Paterson et al. \(1938\)](#) recommended that a scale should be validated by a rigorous item analysis and that only items that correlate highly with the total score should be included. [Hathaway and McKinley \(1940\)](#), the developers of the MMPI, followed this strategy and required that an item only be included on a scale if it discriminated statistically between a criterion group of patients (e.g., those diagnosed with depression) and a sample of individuals from the general population (e.g., visitors to the University of Minnesota Hospitals). Because items are selected based on the prediction of criterion variables, scale content may be heterogeneous. Moreover, empirically derived scales can contain items that overlap with other scales because, in part, the constructs themselves are composed of complex content, not single dimensions.

*Factor-analytic methods.* This approach, often referred to as exploratory factor analysis (e.g., [Cattell 1946; Gorsuch 1963](#)), uses internal statistical methods such as item correlations to develop scales. In this approach, homogeneous item sets are obtained when a pool of items is administered and factor analyzed, with the resulting dimensions then named as scales. Since items for a scale are selected on the basis of item intercorrelations, the scales tend to be homogeneous in content and narrowly defined.

(Continued)

**Table 13.1** (Continued)

*Sequential system of construct-oriented scale development.* Jackson (1970) modified the factor-analytic approach of scale construction. He proposed four principles for test development: (1) Preeminence of psychological theory; (2) Suppression of response style variance; (3) Scale homogeneity and generalizability; (4) Convergent and discriminant validity. This factor-dimensional strategy results in homogeneous content scales that can be recognizable to test takers, thus somewhat prone to response manipulation. More recently, Tellegen et al. (2003) adapted this scale development strategy when constructing the MMPI-2 Restructured Clinical (RC) Scales, but ignored the second principle of suppressing response style variance (Rogers & Sewell, 2006; Rogers et al., 2006). Subsequently, Ben-Porath and Tellegen (2008/2011) replaced the empirically derived MMPI and MMPI-2 Clinical Scales and their empirically derived code types with the RC Scales and other factor-analytically derived scales in the MMPI-2-RF.

*Content grouping with statistical refinement methods.* Another method of constructing personality scales involves grouping items according to similar content as done by Wiggins (1973) following, in part, Cronbach and Meehl's (1955) construct validity approach. Researchers verify their hypothesized item-scale membership using statistical methods like Cronbach's  $\alpha$  coefficient to assure high scale homogeneity. The MMPI-2 Content Scales (Butcher et al. 1989), the MMPI-A Content Scales (Williams, Butcher, Ben-Porath, & Graham, 1992), the Millon Clinical Multiaxial Inventory-III (Millon et al., 2009), and the Personality Assessment Inventory (Morey, 1991) were developed with this method. As with other homogeneous scales that are recognizable to test takers, these scales can be affected by response manipulation.

*Item response theory (IRT).* Although not widely used commercially, personality scales have been developed using IRT. Much of the research using IRT has focused upon scale refinement (Glas, 2009; Thompson, 2009). IRT involves statistically defining characteristics that underlie a particular dimension and statistically evaluating items that match the dimension. It is based on the idea that responses to test items reflect an underlying variable such as a trait (see Embretson, 1996; Reise & Waller, 1993, for more detailed discussion).

*Source:* Adapted with permission from 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 (Butcher, 2010).

McKinley (1940) relied upon the *empirically derived method* to develop the MMPI Clinical Scales, most of the scale construction strategies described in Table 13.1 have been used to develop other MMPI-2 scales in current use, as described later in this chapter.

The MMPI underwent a major revision during the 1980s. The revision included a rewriting of some of the original items to update the language, the deletion of a few items that some found offensive, and the addition of a number of new items to assess problems not addressed in the original MMPI. The MMPI-2 is a 567-item inventory comprised of symptoms, beliefs, and attitudes in adults 18 and older. The revision project also collected a large and more representative normative data set and relevant comparison samples (e.g., psychiatric in- and out-patients, prison inmates, personnel settings) for validity studies of the revised instrument (Butcher,

Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). An updated MMPI-2 manual included new scales and cut-off scores (Butcher et al., 2001).

A separate 467-item version of the instrument, the MMPI-A (Butcher, Williams et al., 1992), was published in 1992. Like the MMPI-2, the MMPI-A contains the original MMPI Validity and Clinical Scales, as well as new scales developed with adolescent samples, designed to assess problems specific to this age group. Readers are referred to Williams and Butcher (2011) for more detailed information about this measure. Both the MMPI-2 and MMPI-A allow for abbreviated administrations. In an abbreviated administration, the respondent completes the first 370 of the MMPI-2 booklet or first 350 items of the MMPI-A booklet (Butcher et al., 2001, 1992). Kruyen et al. (2013) indicated that an abbreviated administration of the MMPI-2 was not a shortened version of the test because it allowed for the scoring of the original Validity Scales and Clinical Scales.

Two other multidimensional inventories of psychopathology, the Millon Clinical Multiaxial Inventories (MCMI; Millon, 1994) and the Personality Assessment Inventory (PAI; Morey, 1991, 2007) have been available for clinical assessments for over 20 years. Both were developed primarily using the *Content Grouping with Statistical Refinement Methods* (Table 13.1). Each put an emphasis on instrument brevity as part of their scale development procedures, resulting in the 175-item MCMI instruments and the 344-item PAI. The MCMI has gone through several versions with the MCMI-III and the MCMI-IV currently available (Millon, Davis, & Millon, 1997; Millon, Millon, Davis, & Grossman, 2009; Millon, Grossman, & Millon, 2015). Additional versions for adolescents and translations into other languages are available for both.

Relatively short personality measures like the Millon inventories are seen by commercial test publishers as having a competitive edge according to Jolosky and Watson (2015). Indeed, Williams and Lally (2017), citing University of Minnesota documents,<sup>1</sup> revealed that marketplace analyses for the MMPI-2 and MMPI-A identified their lengths as weakness. Concerns about the strong competition from the shorter PAI were also highlighted. On the other hand, most surveys of practitioners indicated that use rates of the MMPI/MMPI-2 dwarfed those of the MCMI-III and PAI across a variety of settings (e.g., Archer, Buffington-Vollum, Stredny, & Handel, 2006; Bow, Flens, & Gould, 2010; Camara, Nathan, & Puente, 2000; Fabricatore, Crerand, Wadden, Sarwer, & Krasucki, 2006; Lally, 2003; Rabin et al., 2014; Ready & Veague, 2014; Wise, Streiner, & Walfish, 2010). Several surveys suggest that PAI usage may be increasing (e.g., Lally, 2003; Ready & Veague, 2014; Wise et al., 2010).

In 2002, the University of Minnesota began funding the development of a short form of the MMPI-2 based on the Restructured Clinical (RC) Scales (Tellegen & Ben-Porath, 2002; Williams & Lally, 2017). The RC Scales were introduced in a monograph published by the University in the following year (Tellegen et al.,

<sup>1</sup>The University of Minnesota is the owner and publisher of the MMPI instruments. It licenses Pearson Clinical Assessment as the distributor of the instruments. Its records are subject to public disclosure under the Minnesota Government Data Practices Act.

2003). Tellegen and colleagues (2003) suggested using the RC Scales as aids for the interpretation of the Clinical Scales. The RC Scales monograph (Tellegen et al., 2003) did not describe plans to replace the MMPI-2 Clinical Scales with the RC Scales in a shortened instrument.

The University and Pearson Assessment announced the 338-item MMPI-2-RF at the 2007 convention of the American Psychological Association; it was released the following year (Ben-Porath & Tellegen, 2008/2011; Tellegen & Ben-Porath, 2008/2011). The MMPI-2-RF has been marketed every year since its publication as the standard in MMPI assessments (Isakson, 2007; Lally & Williams, 2017; Williams & Lally, 2017). A 241-item adolescent version, the MMPI-A-RF, was released in 2016, and marketed as the new standard in adolescent personality assessment. However, as Kruyken et al. (2013) indicated, evidence for the validity of an original test does not automatically transfer to a shortened test. Simms, Casillas, Clark, Watson, and Doebling, (2005, p. 357) echoed that point about the foundational scales of the MMPI-2-RF:

*Also, despite the temptation to do so, it is also apparent that the RC Scales cannot be interpreted on the basis of previous empirical studies of the original scales; the RC Scales represent new measures whose meanings must now be determined empirically.*

The scale development procedures used to develop the RC Scales consisted of exploratory factor analysis supplemented by Jackson's (1970) sequential system of construct-oriented scale development (Table 13.1; Rogers & Sewell, 2006; Rogers, Sewell, Harrison, & Jordan, 2006). Yet, Tellegen and colleagues (2003) "entirely ignored one of Jackson's four core principles: the suppression of response-style variance" (Rogers & Sewell, 2006, p. 177). Their efforts were characterized as a "radical retrofitting of the MMPI-2" that "drastically downsized the eight MMPI-2 Clinical Scales by more than 50%, reducing their numbers from 411 to 174 items" (Rogers & Sewell, 2006, p. 175). The 567-item MMPI-2 booklet was downsized to a 338-item measure with 50 scales. Rogers and Sewell (2006) took issue with its developers' advice that the RC Scales could clarify MMPI-2 Clinical Scale findings because that recommendation had neither conceptual nor empirical foundations. Nor did Rogers et al. (2006) agree that the scale construction procedures used to develop the RC Scales were superior to the empirically derived methods or Jackson's sequential system described in Table 13.1.

Other early concerns about the RC Scales were noted in a special issue of the *Journal of Personality Assessment* (e.g., Butcher, Hamilton, Rouse, & Cumella, 2006; Caldwell, 2006; Nichols, 2006). Nichols (2006) pointed to "construct drift" or how fundamentally the constructs measured by RC Scales differed from the MMPI-2 Clinical Scales they ended up replacing on the MMPI-2-RF. Indeed, scale construction procedures for reducing the number of items in a scale that favored selection of highly correlated items tended to narrow construct coverage (Kruyens et al., 2013). Although Archer (2006, p. 179) described the RC Scales as innovative, important, and having "substantial potential," he cautioned that it would be

“premature to discontinue the use of the standard MMPI-2 basic scales until more research has been accumulated on the RC Scales.” Despite these concerns, plans continued to replace the MMPI-2 Clinical Scales with the RC Scales in the MMPI-2-RF, as well as to add new scales introduced for the first time by its authors (Ben-Porath & Tellegen, 2008/2011; Tellegen & Ben-Porath, 2008/2011).

Greene (2011a) called the “MMPI-2” in the name “MMPI-2-RF” a misnomer given how fundamentally different the two instruments were. All six recent textbooks on the MMPI-2 concurred: the MMPI-2-RF was a new test, not a revision of the MMPI-2 (Butcher, Hass, Greene, & Nelson, 2015; Friedman, Bolinskey, Levak, & Nichols, 2015; Graham, 2012; Greene, 2011a; Nichols, 2011). The MMPI-2-RF (Tellegen & Ben-Porath, 2008/2011), and its foundational RC Scales (RC; Tellegen et al., 2003), departed substantially from the MMPI and MMPI-2 scale construction methods, resulting in:

*... a new, and to this point, largely untested psychometric instrument, [which] does not yet have the wealth of empirical support and interpretive data enjoyed by the MMPI-2 (Friedman et al., 2015, p. 593).*

The developers of the RC Scales, the MMPI-2-RF, and their colleagues dismissed these and other critiques, instead providing studies and counterarguments in support of the new instrument and its scales (e.g., Ben-Porath, 2012a, 2012b, 2017; Ben-Porath & Flens, 2012; Tellegen et al., 2006). The publisher and distributor began an extensive marketing campaign promoting the MMPI-2-RF (Isakson, 2007; Lally & Williams, 2017; Williams & Lally, 2017). Greene (2011b, p. 200) warned about the “hyperbole” in the publisher’s statements (e.g., the MMPI-2-RF was the “most significant advancement in clinical personality measurement in decades”). In 2010 the publisher discontinued the scoring program for the widely used 370-item MMPI-2 abbreviated administration. In the year prior to its discontinuation, an estimated 45,527 administrations occurred (Williams & Lally, 2017).

To date, only a few studies surveyed practitioners or clinical psychology training programs about the use of the MMPI-2-RF. Two such studies combined survey responses for the MMPI-2 and MMPI-2-RF, obscuring any differences in usage (Rabin et al., 2014; Wright et al., 2017). In a follow-up to Rabin et al. (2014), Mihura, Roy, and Graceffo (2017) surveyed clinical psychology doctoral training programs, reporting that the MMPI-2 was the most popularly taught adult self-report inventory (92% of programs), followed by the PAI (76%), the MCMI-III (70%), and the MMPI-2-RF (67%). In contrast, Martin, Schroeder, and Odland (2015) used a convenience sample of 316 neuropsychologists to inquire about their use of “symptom-reported validity tests.” They concluded that the MMPI-2-RF was more popular than the MMPI-2 or PAI in this convenience sample and narrow application (Martin et al., 2015). These authors reached a similar conclusion in a subsequent study of responses from 24 neuropsychologists (Schroeder, Martin, & Odland, 2016).

Williams and Lally (2017) proposed an innovative method to assess practitioner use of the MMPI-2, MMPI-2-RF, and MMPI-A. University of Minnesota archival MMPI sales records were examined and used to compare estimated administrations

of the adult instruments. Records were available from 2007 to 2014. They revealed that the MMPI-2 was far more likely to be administered by qualified psychologists than the new test. From 2012 to 2014, more than two-thirds of all administrations of adults were with the MMPI-2, even given the publisher's decision to eliminate scoring products for the abbreviated administration. In a follow-up report, [Lally and Williams \(2017\)](#) reported that preference for the MMPI-2 (61% of all estimated administrations) over the MMPI-2-RF (39%) continued into 2016, even after the introduction of a new MMPI-2-RF scoring report. In contrast, 6 years after the introduction of the MMPI-2, it had achieved near universal professional acceptance (i.e., 92% administrations with the MMPI-2, only 8% with the MMPI). Practitioners' preferred instrument remains the MMPI-2. In the next sections, we describe the various scales used by practitioners to interpret the MMPI-2.

## Assessing protocol validity with the MMPI-2

One of the most valuable aspects of the MMPI-2 is its capability to provide practitioners with information about whether the examinee cooperated with the test administration. Some clients present themselves in ways that their test responses do not represent their true feelings and emotions. For example, if a person is being evaluated to determine their mental health status for a pretrial criminal evaluation, they might exaggerate their symptoms to appear psychologically disturbed. Or, if they are applying for high-risk positions such as an airline pilot or police officer, they may present themselves as highly virtuous and free of mental health or personality problems. [Table 13.2](#) presents the 11 measures on the MMPI-2 that provide information about the examinee's response approach.

Intentionally missing from [Table 13.2](#) is the Symptom Validity Scale, originally named the Fake Bad Scale (FBS; [Lees-Haley, English, & Glenn, 1991](#)). The FBS was introduced as an MMPI-2 scale designed to detect malingering in personal injury claimants. A shortened version of the 43-item FBS, the FBS-R (30 items) was included on the MMPI-2-RF ([Ben-Porath & Tellegen, 2008/2011](#)). [Lees-Haley et al. \(1991\)](#) used a *rational or content-based* method when selecting items for the FBS ([Table 13.1](#)). Three years earlier, [Lees-Haley \(1988\)](#) also employed the *rational or content-based* item-selection strategy and used a majority of the items appearing on the FBS and FBS-R to develop a scale measuring an entirely different construct that he called the Litigation Response Syndrome (LRS; [Nichols & Gass, 2015](#)). According to Lees-Haley (1988, p. 110):

*Litigation is often so troublesome that it leads to Litigation Response Syndrome, known as LRS, which is a group of stress problems caused by the process of litigation. LRS is made up of complaints that arise solely from the experience of being personally involved in a lawsuit, rather than from the events that precipitated the litigation. It is an ephemeral consequence of litigation, not a permanent problem, and needs to be distinguished from allegedly enduring problems it resembles.*

**Table 13.2** The MMPI-2 validity scales

Scale name	Scale description
? (Cannot Say)	The number of items that are left blank or are answered in both directions. An invalid profile is suggested if the ? raw score is 30 or more.
All True	The respondent does not follow instructions and marks all items as “True.” A highly elevated invalid profile results.
All False	The respondent does not follow instructions and marks all items as “False.” A distinctive invalid profile results.
VRIN (Variable Response Inconsistency Scale)	The VRIN scale consists of pairs of items that have either similar or opposite content; each pair is scored for the occurrence of an inconsistency in responses to the two items. A high VRIN score is a warning that a test subject may have answered the items in an indiscriminate manner.
TRIN (True Response Inconsistency Scale)	The TRIN scale consists of 20 pairs of items that are opposite in content. If a subject responds inconsistently by answering True to both items of certain pairs, one point is added to the TRIN score. High TRIN scores indicate an invalid profile due to indiscriminate True responding.
F (Infrequency)	This scale contains items that are infrequently endorsed by most people. A high score suggests an exaggerated of symptoms that is inconsistent with accurate self-appraisal.
F(B) (Infrequency-Back)	The F(B) scale is comprised of infrequent items contained in the back half of the MMPI-2 booklet. Its interpretation is similar to the F scale.
F(P) (Psychopathology Infrequency)	Measures infrequent psychopathology or responding to an inordinate number of extreme items compared to a psychiatric sample.
L (Lie)	The L scale addresses a somewhat unsophisticated or “virtuous” test-taking attitude. Elevated scores suggest an individual who presents an overly positive picture, creating an unrealistically favorable view of his or her adjustment.
K (Defensiveness)	High scores reflect an uncooperative attitude and an unwillingness or reluctance to disclose personal information or problems. Low scores suggest openness and frankness. K is positively correlated with intelligence and educational level, which should be considered.
S (Superlative Self-Presentation Scale)	The S scale measures test defensiveness and overly positive presentation.

Paradoxically, Lees-Haley (1991) selected 23 items from the LRS to include on the FBS. An additional 12 other items that are a demand characteristic associated with plaintiff status were also included on the FBS. These items include denial of occasional dishonesty and other virtue-related items that, like the LRS items, are inconsistent with a malingering or faking bad response style (Gass & Odland, 2012, 2014; Nichols & Gass, 2015).

Despite these obvious flaws, the University of Minnesota Press added the FBS to its scoring programs for the MMPI-2 in 2007. The next year the FBS-R was included in the MMPI-2-RF. The MMPI publisher continues to promote use of the FBS despite multiple critiques in the peer-reviewed literature and numerous court challenges to its scientific credibility leading to its inadmissibility in expert witness testimony (e.g., Butcher, Gass, Cumella, Kally, & Williams, 2008; Gass, Williams, Cumella, Butcher, & Kally, 2010; Nichols & Gass, 2015; Williams, Butcher, Gass, Cumella, & Kally, 2009). For more information about the problems with the FBS the reader is referred to Butcher et al. (2015) and Friedman et al. (2015).

## Assessing psychopathology: MMPI-2 clinical scales

The mainstay of the MMPI and MMPI-2 are the empirically derived Clinical Scales: Scale 1 Hypochondriasis (*Hs*); Scale 2, Depression (*D*); Scale 3, Hysteria (*Hy*); Scale 4, Psychopathic Deviation (*Pd*); Scale 6, Paranoia (*Pa*); Scale 7, Psychastenia (*Pt*); Scale 8, Schizophrenia (*Sc*); and Scale 9, Mania (*Ma*). In addition, two other scales were added to the clinical profile: Scale 5 Masculinity/Femininity (*Mf*) and Social Introversion (*Si*). If a single scale is elevated on a given profile (i.e., *T-score* greater than 65), then the empirically based correlates for that scale are applied as descriptors for a given patient. A number of recent texts provide descriptors for the MMPI-2 Clinical Scales (e.g., Butcher, 2011; Greene, 2011a; Friedman et al., 2015). Table 13.3 provides a summary of empirically validated descriptors for these scales.

Very early in the development of the MMPI, researchers discovered that many patients had significant elevations on more than one of the Clinical Scales. Furthermore, patients with many two- and three-scale elevations showed similar behaviors and symptoms. These different patterns were referred to as code types. For example, elevations on the Depression and Psychastenia Scales were referred to as the 2–7/7–2 code type (the scale score with the highest elevation is listed first in the code type). Empirically established descriptors for the 2–7/7–2 code type include symptoms of depression, anxiety and tension. Recent MMPI-2 texts summarize the various code types and their empirical descriptors (e.g., Butcher, 2011; Greene, 2011a; Friedman et al., 2015).

**Table 13.3** The MMPI-2 clinical scales

Scale name	Empirically derived descriptors
1 (Hs, Hypochondriasis)	High scorers present multiple, vague, and chronic physical problems. They are likely viewed as unhappy, self-centered, hostile, and attention seeking.
2 (D, Depression)	High scorers present with depressed mood, low self-esteem, lethargy, and feelings of guilt.
3 (Hy, Hysteria)	Neurotic defenses (e.g., denial and repression) are common. They tend to develop physical symptoms in reaction to stress and can be dependent, naïve, infantile, and narcissistic.
4 (Pd, Psychopathic Deviate)	High scores are associated with antisocial behavior (e.g., rebelliousness, family problems, impulsivity, legal problems, and alcohol or drug abuse).
5 (Mf, Masculinity/Femininity)	High scoring men may have interests more traditionally viewed as feminine. High scoring women may be more interested in traditionally masculine pursuits.
6 (Pa, Paranoia)	Scale elevations are often associated with being suspicious, aloof, guarded, and overly sensitive. High scorers may externalize blame and hold grudges.
7 (Pt, Psychasthenia)	High scorers are tense, anxious, ruminative, obsessional, phobic, and rigid. They frequently are self-condemning and guilt prone, and resist psychological interpretations.
8 (Sc, Schizophrenia)	High scorers may be withdrawn, moody, and confused. Unusual thoughts, poor judgment, and erratic behavior can be present. Bizarre sensory experiences, delusions, and hallucinations are more likely with extreme scores.
9 (Ma, Hypomania or Mania)	High scorers are viewed as sociable and optimistic, although they can be manipulative and grandiose. Affective disorders, bizarre behavior, erratic mood, impulsivity, and delusions may be present in those with very high scores.
0 (Si, Social Introversion)	High scorers are likely to be introverted, withdrawn, submissive, over-controlled, tense, inflexible, and guilt prone. Low scorers can be gregarious, expressive, talkative, impulsive, uninhibited, manipulative, and possibly insincere in social relations.

## Assessing psychopathology: MMPI-2 supplementary scales

Over the years, researchers developed additional scales (i.e., the Supplementary Scales) found to provide information beyond the original Clinical Scales. Three of the Supplementary Scales measure problems related to alcohol and drug use ([MacAndrew, 1965](#); [Weed, Butcher, McKenna, & Ben-Porath, 1992](#)), one assesses

**Table 13.4** The MMPI-2 supplementary scales

Scale name	Scale description
MAC-R (MacAndrew Alcoholism Scale)	MAC-R Scale is an empirically derived scale that distinguished alcoholic from nonalcoholic psychiatric patients. It is a measure of the proneness to developing or having an addiction problem, including drug abuse or pathological gambling.
APS (Addiction Potential Scale)	The MMPI-2 item pool, which included a broader range of alcohol and drug items than did the original MMPI, was used to empirically derive the APS. Only 9 of its 39 items overlap with MAC-R. High scores are associated with greater likelihood of developing problems with alcohol and other drugs.
AAS (Addiction Acknowledgment Scale)	AAS was developed using the rational or content-based item selection method that was followed-up with statistical verification of item-scale membership. This resulted in a scale, unlike MAC-R and APS, with face valid items (i.e., those asking about alcohol or other drug use behaviors). High scorers acknowledge having a large number of alcohol or drug problems in comparison to the normative sample.
MDS (Marital Distress Scale)	MDS is an empirically derived measure of marital distress and relationship problems. It is only appropriate for use with clients who are married or separated.
Pk (Post-Traumatic Stress Disorder Scale)	Pk is a measure of post-traumatic stress disorder (PTSD) developed using an empirical scale construction strategy contrasting a sample of 100 male veterans diagnosed with PTSD with 100 male veterans with other psychiatric problems.

clients' attitudes toward their marital relationship (Hjemboe, Almagor & Butcher, 1992) and one measures an individual's reaction to trauma (Keane, Malloy, & Fairbank, 1984). Table 13.4 describes the Supplementary Scales.

## Assessing psychopathology: MMPI-2 content scales

Early MMPI research and clinical applications emphasized empirical correlates and objective interpretation of code types. Content interpretation assumes that when clients respond validly to MMPI items, they are responding directly to item content and meaning. Consequently, individual's responses to the content of the MMPI item pool could provide information not available through empirical test interpretation procedures. Table 13.1 includes several scale construction methods used to develop content scales. The 15 MMPI-2 Content Scales were developed using the

**Table 13.5** The MMPI-2 content scales

Scale name	Interpretive statements for high scorers
ANX (Anxiety)	Reports feeling anxious, insecure, indecisive, apprehensive and having concentration problems or sleep difficulties.
FRS (Fears)	Reports fears, possible phobias; uneasy, low self-confidence.
OBS (Obsessiveness)	Problems making decisions, rigid thinking, dislikes change, ruminations. May feel depressed and lacks self-confidence.
DEP (Depression)	Depressed, despondent, pessimistic, fatigued, lacks interest in things, hopeless, and guilty. Possible suicidal ideation.
HEA (Health Concerns)	Reports poor physical health, preoccupation with bodily functions, may develop physical symptoms when under stress.
BIZ (Bizarre Mentation)	Endorses psychotic symptoms (e.g., hallucinations, delusions, paranoid ideations). Possibly disoriented.
ANG (Anger)	Anger-control problems, irritability, impatience, loss of control, temper tantrums, past or potential abusiveness.
CYN (Cynicism)	Sees others as selfish, dishonest, and uncaring. Suspicious of others. Hostile, guarded, and untrusting in relationships.
ASP (Antisocial Practices)	Reports antisocial behaviors, cynical attitudes, and resentment of authority. Likely to be manipulative, self-centered, aggressive, and impulsive. May have alcohol/drug problems.
TPA (Type A)	Hard-driving, work-oriented behaviors. Feelings of time pressure, impatience, irritability, hostility, and easily annoyed.
LSE (Low Self-Esteem)	Reports feeling inept, unattractive, unlikeable, and useless.
SOD (Social Discomfort)	Feels socially awkward, dislikes large gatherings, is uneasy around others, and prefers being alone.
FAM (Family Problems)	Feels angry and hostile toward family members; may have negative views about marriage; describes their family as lacking in love; does not view family as a source of support.
WRK (Work Interference)	Negative attitudes toward career or coworkers; poor decision making; lacks ambition or energy. Also, low self-esteem, obsessiveness, and tension.
TRT (Negative Treatment Indicators)	Feels that no one can understand them; beliefs that change is not possible; poor problem solvers; guarded; negative attitudes toward doctors or mental health treatment.

*Content Grouping with Statistical Refinement Method* (Butcher, Graham, Williams, & Ben-Porath, 1990). In some cases, the MMPI-2 Content Scale (e.g., Depression, DEP) is an updated version of a MMPI Wiggins Content Scale (Wiggins, 1973). In other instances, these scales are constructs based on item content added to the MMPI-2 and not previously identified on the MMPI (e.g., Type A [TPA]; Work Interference [WRK]; Negative Treatment Indicators [TRT]). Table 13.5 describes the MMPI-2 Content Scales.

## Publications on the MMPI-2, MMPI-2-RF, PAI, and MCMI

The previous edition of this *Handbook* provided information about the number of publications on the MMPI, MCMI, and PAI for the years 1990–94 (Nezami & Butcher, 1999). During that period, the MMPI was by far the most widely researched instrument for personality assessment. It showed a steady increase from 149 publications in 1990 to 277 in 1994. During the same years the peak number of MCMI publications was 46. Publications for the more recently released PAI ranged from 3 to 9 between 1990 and 1994, with the largest number in 1991.

An updated search using the American Psychological Association's (APA) PsycINFO database of scholarly publications on the MMPI-2, MMPI-2-RF, the PAI, and the MCMI was completed in July 2017 (Table 13.6). All publications from 2003 to 2016 were included in Table 13.6 with the exceptions of book reviews, errata, updates or republications, and publications with only brief remarks about the instrument (e.g., in a single sentence or use of a few items from the instrument). Authored books counted as one publication, whereas chapters by different authors in an edited book were counted individually.

**Table 13.6** PsycINFO publications on the MMPI-2, MMPI-2-RF, PAI, and MCMI (2003–2016)

	MMPI/MMPI-2 (overlap/% MMPI-2-RF) <sup>a</sup>	MMPI-2-RF (overlap/% RC Scales) <sup>b</sup>	PAI (overlap/% MMPI) <sup>c</sup>	MCMI (overlap/% MMPI)	Total (without overlap)
2016	64 (07/11%)	38 (0)	75 (05/07%)	15 (01/07%)	179
2015	47 (12/26%)	45 (0)	80 (02/03%)	17 (02/12%)	173
2014	69 (02/03%)	39 (0)	67 (03/04%)	18 (03/17%)	185
2013	105 (17/16%)	54 (04/07%)	104 (01/01%)	26 (03/12%)	268
2012	91 (17/19%)	37 (05/14%)	91 (01/01%)	31 (05/16%)	227
2011	103 (15/15%)	32 (04/13%)	59 (02/03%)	24 (04/17%)	197
2010	117 (08/07%)	19 (04/21%)	84 (04/05%)	28 (03/11%)	233
2009	120 (07/06%)	09 (07/78%)	68 (01/01%)	28 (02/07%)	215
2008	154 (19/12%)	21 (19/90%)	58 (03/05%)	52 (09/17%)	254
2007	122 (10/08%)	10 (10/100%)	56 (07/13%)	33 (01/03%)	203
2006	161 (11/07%)	11 (11/100%)	42 (05/12%)	26 (06/23%)	218
2005	136 (06/04%)	06 (06/100%)	40 (04/10%)	38 (04/11%)	206
2004	168 (0)	00 (0)	31 (00/00%)	40 (07/18%)	232
2003	189 (1/005%)	01 (01/100%) <sup>d</sup>	25 (03/12%)	35 (09/26%)	237
Total	1646	322	880	411	

<sup>a</sup>The numbers in parentheses indicate the publications that examined both the MMPI-2 and the MMPI-2-RF or its RC Scales (e.g., in 2009 there were 120 publications on the MMPI-2, of which seven, or 6%, also examined the MMPI-2-RF and/or the RC Scales).

<sup>b</sup>RC Scales studies are included with the MMPI-2-RF. The overlap of studies exclusively on the RC Scales with MMPI-2-RF studies are in parentheses (e.g., of the nine publications in 2009, seven—or 78%—were RC Scales only).

<sup>c</sup>The PAI and MCMI columns include information on the overlap of these instruments with any of the MMPI instruments (e.g., in 2009, only one or 1% of the 68 PAI publications also examined an MMPI instrument; and only two or 7% of the 28 MCMI publications included an MMPI instrument).

<sup>d</sup>This publication is the RC Scales monograph, published by the University of Minnesota Press. It was not indexed in PsycInfo.

Broad search terms (e.g., MMPI, MCMI) were used. MMPI and MCMI were relatively unique acronyms in PsycINFO, and returned results for all the instruments (e.g., MMPI, MMPI-2-RF, MCMI, MCMI-III). Date of publication was the only filter used in these searches. “Personality Assessment Inventory” and “PAI” were more common character strings in PsycINFO, yielding thousands of results, many irrelevant (e.g., the surname, PAI, or the human protein PAI-1). In addition to year, results for the PAI were filtered using the “Tests and Measures” option.

Each search result was examined to determine which instrument (or combination of instruments) was the focus. The categories in [Table 13.6](#) are not mutually exclusive. If, for example, a publication reported both MMPI-2 and MMPI-2-RF data, it was counted in both categories. Prior to the publication of the MMPI-2-RF, studies on the RC Scales were included as both MMPI-2 and MMPI-2-RF publications. Although the period covered in [Table 13.6](#) begins 3 years after the discontinuation of sales of the MMPI, some publications used the term MMPI. In the MMPI-2 category, the percentage of publications using the term MMPI ranged from 12% to 30% ( $M = 19.92\%$ ,  $SD = 5.25\%$ ). Overlap between any MMPI instrument and the PAI or MCMI was counted similarly. Overlap of the PAI with the MCMI was rare and not included in [Table 13.6](#). With the exception of the overlap of the PAI with the MCMI, the total column for each year eliminates overlap. Because the PAI and MCMI overlap was not included in [Table 13.6](#), there may be a slight overestimate of the number of publications in a given year.

[Table 13.6](#) reveals the continued domination of the MMPI-2 in personality assessment research from 2003 to 2016 with its total of 1646 publications. The PAI is the second most widely researched inventory during this period with 880 publications, followed by the MCMI with 411, and the MMPI-2-RF with 322. Publications on the MCMI in [Table 13.6](#) (range 15–52) are fairly consistent with what was reported in the last edition of this *Handbook* (range 25–46). Research on the PAI demonstrates a steady increase since its introduction in 1991. Not unexpectedly for a new instrument, the MMPI-2-RF, published in 2008 with its core RC Scales introduced in 2003, shows a clear pattern of increased publications from 2003 to 2016; the MMPI-2 demonstrates the opposite trend with its 189 publications in 2003 dropping to 64 in 2015.

Supporters of the MMPI-2-RF argue that the increase in MMPI-2-RF publications since its introduction demonstrates the general acceptance in the field of this new instrument. For example, the publisher’s website currently features interviews with forensic and MMPI-2-RF experts who clearly favor use of the MMPI-2-RF over the MMPI-2. [Boone \(2017\)](#) suggests that psychologists using the MMPI-2, rather than the MMPI-2-RF, in court testimony are more likely to have to justify their choice, in part because “there are nearly 300 peer-reviewed publications on the MMPI-2-RF.” Similarly, [Sellbom \(2017\)](#) points to how impressive it is that “the MMPI-2-RF scales have appeared in almost 300 peer-reviewed articles . . . given how long these scales have actually been available.” Earlier [Ben-Porath \(2012a, p. 691\)](#) recommended that challenges to expert witness testimony based on the MMPI-2-RF could be addressed

by referring to the MMPI-2-RF test documents (i.e., Ben-Porath & Tellegen, 2008/2011; Tellegen & Ben-Porath, 2008/2011) and the “over 150 articles published in peer-reviewed journals” he had identified as of August 2012.

However, the “nearly 300” MMPI-2-RF publications include many that highlighted problems with the RC Scales, the scale development approach taken, or the need for more research before using them to refine MMPI-2 interpretations, let alone to form the core of a new test (see pp. 7–10 above for details; and Butcher et al., 2015; Friedman et al., 2015; Graham, 2012; Greene, 2011a; Nichols, 2011). Prior to the release of the MMPI-2-RF, the 28 articles in the MMPI-2-RF column from 2003 to 2007 were exclusively about the RC Scales (i.e., no research was presented on the other 42 MMPI-2-RF scales until publication of its manuals the following year). Furthermore, simply reporting the existence of 300+ studies on the MMPI-2-RF without discussing comparable figures for the MMPI-2 (over five times as many studies in the same time period), the PAI (almost three times as many), or the MCMI (almost 100 more studies) provides an incomplete picture about the adequacy of the research base supporting the MMPI-2-RF.

## Concluding comments

Of the four multidimensional self-report inventories of psychopathology, the MMPI-2 remains by far the most widely researched instrument. Despite significant investment in marketing, research, and development by its publisher and distributor, the MMPI-2-RF lags behind the other three multidimensional inventories in terms of research publications (Table 13.6). It also lags by a three to two margin behind the MMPI-2 in number of administrations by qualified practitioners (Lally & Williams, 2017). Indeed, use of the current adult versions of the MMPI (i.e., either the MMPI-2 or MMPI-2-RF) has declined since the introduction of the new instrument and the ensuing controversy (Lally & Williams, 2017; Williams & Lally, 2017). A topic for further research is the effect of the ubiquitous marketing taglines suggesting that the MMPI-2-RF is the current standard in personality assessment, the publisher’s preferential discretionary funding of MMPI-2-RF research over that on the MMPI-2, and its elimination of MMPI-2 products (e.g., scoring for the abbreviated administration) on the declines in MMPI-2 use and publications.

Research on the PAI increased during this period. Currently, no information is available to determine if psychologists switched from the MMPI instruments to another multidimensional inventory like the PAI, single dimension brief measures like the Beck inventories, or reduced their use of testing. Surveys of practitioners can address this issue. Finally, we agree with others who indicate that it is time for psychologists to pay more attention to marketing’s impact on assessment, to more carefully consider the push for shorter tests, and to develop more rigorous standards for new psychological tests prior to their marketing and distribution (Greene, 2011b; Kruyken et al., 2012; Kruyken et al., 2013; Loring & Bauer, 2010).

## References

- Anastasi, A. (1982). *Psychological testing* (5th ed). New York, NY: Macmillan.
- Archer, R. P. (2006). A perspective on the Restructured Clinical Scales (RC) scale project. *Journal of Personality Assessment*, 87, 179–185.
- 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(1), 84–94.
- Beck, A. T., & Steer, R. A. (1990). *Manual for the Beck Anxiety Inventory*. San Antonio, TX: Psychological Corporation.
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.
- Ben-Porath, Y. S. (2012a). Addressing challenges to MMPI-2-RF-based testimony: Questions and answers. *Archives of Clinical Neuropsychology*, 27, 691–705.
- Ben-Porath, Y. S. (2012b). *Interpreting the MMPI-2-RF*. Minneapolis, MN: University of Minnesota Press.
- Ben-Porath, Y. S. (2017). An update to Williams and Lally's (2017) analysis of MMPI-2-RF acceptance. *Professional Psychology: Research and Practice*, 48, 275–278. Available from <https://doi.org/10.1037/pro0000115>.
- Ben-Porath, Y. S., & Flens, J. R. (2012). Butcher and Williams's (this issue) critique of the MMPI-2-RF is slanted and misleading. *Journal of Child Custody*, 9, 223–232.
- Ben-Porath, Y. S., & Tellegen, A. (2008/2011). *MMPI-2-RF manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.
- Boone, K. (July 17, 2017). Interview with a forensic expert [Publisher's webpage]. Retrieved from <<http://www.upress.umn.edu/test-division/MMPI-2-RF/mmpi-2-rf-expert-interviews#Boone>>.
- 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.
- 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.
- Butcher, J. N. (2011). *The MMPI-2: A beginner's guide* (3rd ed). Washington, DC: American Psychological Association.
- 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, MN: 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 the Law*, 1, 191–209.
- Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2: Manual for administration, and scoring*. Minneapolis, MN: 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, MN: 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(1), 199–205.
- Butcher, J. N., Hass, G. A., Greene, R. L., & Nelson, L. D. (2015). *Using the MMPI-2 in forensic assessment*. Washington, DC: American Psychological Association.

- 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, MN: University of Minnesota Press.
- Caldwell, A. B. (2006). Maximal measurement or meaningful measurement: The interpretive challenges of the MMPI-2 Restructured Clinical (RC) Scales. *Journal of Personality Assessment*, 87, 193–201.
- 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.
- Cattell, J. M. (1890). Mental tests and measurements. *Mind*, 15, 373–381.
- Cattell, R. B. (1946). *Description and measurement of personality*. Yonkers-on-Hudson, NY: World Book Company.
- Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (1993). *Sixteen personality factor questionnaire* (5th ed). Champaign, IL: Institute for Personality and Ability Testing.
- 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.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 2, 281–302.
- Embreton, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8, 341–349.
- Fabricatore, A. N., Crerand, C. E., Wadden, T. A., Sarwer, D. B., & Krasucki, J. L. (2006). How do mental health candidates evaluate candidates for bariatric surgery? *Obesity Surgery*, 16, 567–573.
- Friedman, A. F., Bolinskey, P. K., Levak, R., & Nichols, D. S. (2015). *Psychological assessment with the MMPI-2/RF* (3rd ed). New York: Routledge/Taylor & Francis.
- Galton, F. (1884). Measurement of character. *Fortnightly Review*, 42, 179–185.
- Gass, C. S., & Odland, A. P. (2012). Minnesota Multiphasic Personality Inventory-2 Restructured Form Symptom Validity Scale-Revised (MMPI-2-RF FBS-r; also known as Fake Bad Scale): Psychometric characteristics in a nonlitigation neuropsychological setting. *Journal of Clinical and Experimental Neuropsychology*, 34(6), 561–570. Available from <https://doi.org/10.1080/13803395.2012.666228>.
- Gass, C. S., & Odland, A. P. (2014). The MMPI-2 Fake Bad Scale (FBS): Psychometric characteristics and limitations in a non-litigation neuropsychological setting. *Applied Neuropsychology: Adult*, 21(1), 1–8.
- 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*, 3, 81–85.
- Gorsuch, R. L. (1963). *Factor analysis* (2nd ed). Hillsdale, NJ: LEA Press.
- Gough, H. G. (1956). *California psychological inventory*. Palo Alto, CA: Consulting Psychologists Press.
- Glas, C. A. W. (2009). What IRT can and cannot do. *Measurement: Interdisciplinary Research and Perspectives*, 7, 91–93.
- Graham, J. R. (2012). *MMPI-2: Assessing personality and psychopathology* (5th ed). New York, NY: Oxford University Press.

- Greene, R. L. (2011a). *MMPI-2/MMPI-2-RF: An interpretive manual* (3rd ed). Boston, MA: Allyn & Bacon.
- Greene, R. L. (2011b). Some considerations for enhancing psychological assessment. *Journal of Personality Assessment*, 93, 198–203.
- Hathaway, S. R. (1965). Personality inventories. In B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 451–476). New York, NY: McGraw-Hill.
- Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): 1. Construction of the schedule. *Journal of Psychology*, 10, 249–254.
- Heymans, G., & Wiersma, E. (1906). Beitrage zur spezilien psychologie auf grund einer massenunterschung. *Zeitschrift Fur Psychologie*, 43, 81–127.
- Hjemboe, S., Almagor, M., & Butcher, J. N. (1992). Empirical assessment of marital distress: The Marital Distress Scale (MDS) for the MMPI-2. In C. D. Spielberger, & J. N. Butcher (Eds.), *Advances in personality assessment* (Vol. 9, pp. 141–152). New Jersey: Lawrence Erlbaum Press.
- Isakson, K. (2007). *MMPI-2, MMPI-2-RF, and MMPI-A marketing plan overview*. Minneapolis, MN: Pearson, [University of Minnesota Documents Nos. 341-347 provided under a 2009 Minnesota Government Data Practices Request].
- Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2, pp. 60–96). New York, NY: Academic.
- Jolosky, T., & Watson, C. (2015). The business of testing. *Journal of Personality Assessment*, 97, 597–604.
- Keane, T. M., Malloy, P. F., & Fairbank, J. A. (1984). Empirical development of an MMPI subscale for the assessment of posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, 52, 888–891.
- Kruyten, P. M., Emmons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, 12, 321–344. Available from <https://doi.org/10.1080/15305058.2011.643517>.
- Kruyten, P. M., Emmons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, 13, 223–248. Available from <https://doi.org/10.1080/15305058.2012.703734>.
- Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research and Practice*, 34, 491–498.
- Lally, S. J., & Williams, C. L. (2017). Response to Ben-Porath's update to Williams and Lally (2016). *Professional Psychology: Research and Practice*, 48, 282–285. Available from <https://doi.org/10.1037/pro0000157>.
- Lees-Haley, P. R. (1988). Litigation response syndrome. *American Journal of Forensic Psychology*, VI, 3–12.
- 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.
- Loring, D. W., & Bauer, R. M. (2010). Testing the limits: Cautions and concerns regarding the new Wechsler IQ and Memory scales. *Neurology*, 74, 685–690.
- MacAndrew, C. (1965). The differentiation of male alcoholic outpatients from nonalcoholic psychiatric outpatients by means of the MMPI. *Quarterly Journal of Studies on Alcohol*, 26, 238–246.
- Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey of North American professionals. *The Clinical Neuropsychologist*, 29(6), 741–776. Available from <https://doi.org/10.1080/13854046.2015.1087597>.

- Meehl, P. E. (1945). The dynamics of “structured” personality tests. *Journal of Clinical Psychology*, 1, 296–303.
- Mihura, J. L., Roy, M., & Graceffo, R. A. (2016). Psychological assessment training in clinical psychology doctoral programs. *Journal of Personality Assessment*, 99, 153–164. Available from <https://doi.org/10.1080/00223891.2016.1201978>.
- Millon, T. (1994). *The Millon Clinical Multiaxial Inventory-III manual*. Minneapolis, MN: National Computer Systems.
- Millon, T., Davis, R., & Millon, C. (1997). *The Millon Clinical Multiaxial Inventory-III manual* (2nd ed.). Minneapolis, MN: National Computer Systems.
- Millon, T., Grossman, S., & Millon, C. (2015). *MCM-IV manual* (4th ed.). Minneapolis, MN: Pearson Clinical.
- Millon, T., Millon, C., Davis, R., & Grossman, S. (2009). *MCM-III manual* (3rd ed.). Minneapolis, MN: Pearson Clinical.
- Morey, L. C. (1991). *Personality assessment inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Morey, L. C. (2007). *Personality assessment inventory: Professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.
- Nezami, E., & Butcher, J. N. (1999). Objective personality assessment. In G. Goldstein, & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 413–436). New York, NY: Pergamon Press.
- Nichols, D. S. (2006). The trials of separating bath water from baby: A review and critique of the MMPI<sup>2</sup> Restructured Clinical Scales. *Journal of Personality Assessment*, 87, 121–138.
- Nichols, D. S. (2011). *Essentials of MMPI-2 assessment* (2nd ed.). New York, NY: John Wiley & Son.
- Nichols, D. S., & Gass, C. S. (2015). The Fake Bad Scale: Malingering or litigation response syndrome—which is it? *Archives of Assessment Psychology*, 5, 5–10.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
- Paterson, D. G., Schneidler, G. G., & Williamson, E. G. (1938). *Student guidance techniques: A handbook for counselors in high schools and colleges*. New York, NY: McGraw-Hill Book Company.
- Rabin, L. A., Spadaccini, A. T., Brodale, D. L., Grant, K. S., Elbulok-Charcape, M. M., & Barr, W. B. (2014). Utilization rates of computerized tests and test batteries among clinical neuropsychologists in the United States and Canada. *Professional Psychology: Research and Practice*, 45(5), 368–377.
- Ready, R. E., & Veague, H. B. (2014). Training in psychological assessment: Current practices of clinical psychology programs. *Professional Psychology: Research and Practice*, 45(4), 278–282. Available from <https://doi.org/10.1037/a0037439>.
- Reise, S. P., & Waller, N. (1993). Traitness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, 65, 143–151.
- Rogers, R., & Sewell, K. W. (2006). MMPI-2 at the crossroads: Aging technology or radical retrofitting. *Journal of Personality Assessment*, 87(2), 175–178.
- 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 (2), 139–147. Available from [https://doi.org/10.1207/s15327752jpa8702\\_03](https://doi.org/10.1207/s15327752jpa8702_03).
- Schroeder, R. W., Martin, P. K., & Odland, A. P. (2016). Expert beliefs and practices regarding neuropsychological validity testing. *The Clinical Neuropsychologist*, 30, 515–535. Available from <https://doi.org/10.1080/13854046.2016.1177118>.

- Sellbom, M. (July 17, 2017). Interview with an MMPI-2-RF expert [Publisher's webpage]. Retrieved from <<http://www.upress.umn.edu/test-division/MMPI-2-RF/mmpi-2-rf-expert-interviews>>.
- 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.
- Tellegen, A., & Ben-Porath, Y. S. (April 15, 2002). *Exploration of developing an MMPI-2 short form* [University of Minnesota Documents provided under a 2007 Minnesota Government Data Practices Request.]
- Tellegen, A., & Ben-Porath, Y. S. (2008/2011). *MMPI-2-RF technical manual*. Minneapolis, MN: University of Minnesota Press.
- Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The MMPI-2 restructured clinical scales: Development, validation, and interpretation*. Minneapolis, MN: University of Minnesota Press.
- Tellegen, A., Ben-Porath, Y. S., Sellbom, M., Arbisi, P. A., McNulty, J. L., & Graham, J. R. (2006). Further evidence on the validity of the MMPI-2 restructured clinical (RC) scales: Addressing questions raised by Rogers, Sewell, Harrison, and Jordan and Nichols. *Journal of Personality Assessment, 87*, 148–171.
- Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778–793.
- Weed, N. C., Butcher, J. N., McKenna, T., & Ben-Porath, Y. S. (1992). New measures for assessing alcohol and drug abuse: The APS and AAS. *Journal of Personality Assessment, 58*, 389–404.
- Wiggins, J. S. (1973). *Personality and prediction*. Reading, MA: Addison-Wesley.
- Williams, C. L., & Butcher, J. N. (2011). *A beginner's guide to the MMPI-A*. Washington, DC: American Psychological Association.
- Williams, C. L., Butcher, J. N., Ben-Porath, Y. S., & Graham, J. R. (1992). *MMPI-A Content Scales: Assessing psychopathology in adolescents*. Minneapolis: University of Minnesota Press.
- 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, & Kaufman (2009). *Psychological Injury and the Law, 2*, 182–197.
- Williams, C. L., & Lally, S. (2017). MMPI-2, MMPI-2-RF, and MMPI-A administrations (2007–2014): Any evidence of a “new standard?” *Professional Psychology: Research and Practice, 48*, 267–274. Available from <https://doi.org/10.1037/pro0000088>.
- Wise, E. A., Streiner, D. L., & Walfish, S. (2010). A review and comparison of the reliabilities of the MMPI-2, MCMI-III, and PAI presented in their respective test manuals. *Measurement and Evaluation in Counseling and Development, 42*, 246–254.
- Woodworth, R. S. (1919). Examination of emotional fitness for war. *Psychological Bulletin, 15*, 59–60.
- Woodworth, R. S. (1920). *Personal data sheet*. Chicago, IL: Stoelting.
- Wright, C. V., Bettie, S. G., Galper, D. I., Church, A. S., Bufka, L. F., Brabender, V. M., ... Smith, B. L. (2017). Assessment practices of professional psychologists: Results of a national survey. *Professional Psychology: Research and Practice, 48*, 73–78.

# The Rorschach

14

*Philip Erdberg*

University of California - San Francisco, CA, United States

## Introduction

The assessment technique that Hermann Rorschach introduced in 1921 has certainly had its share of critics over the decades. But even they must concede the resilience of an instrument that, against considerable odds, will celebrate a healthy 100th anniversary in just a few years.

Rorschach died just a year after the test's initial publication, leaving the fledgling instrument in the hands of three of his associates. In little more than a decade it had traveled to America, where it found itself with five groups of increasingly diverging adoptive parents. Methodological criticism came from outside the Rorschach community as well, and there were suggestions that the test be discarded entirely.

But by the mid-1970s a consolidation integrated the best of what had been learned during the half century of divergence, and early in the next century a new system took advantage of the increasing accumulation of research to bring the test in line with contemporary standards. The history of the test's development and a description of current research and clinical trends are the subjects of this chapter.

## History and development

The idea that associations to ambiguous visual stimuli could lead to understanding a person is an ancient one. The classical Greeks were interested in the interaction of ambiguity and the person's representation of reality (Piotrowski, 1957). And by the 15th century, both DaVinci and Botticelli had postulated a relationship between creativity and the processing of ambiguous materials (Zubin, Eron, & Schumer, 1965). Using inkblots as stimuli for imagination achieved substantial popularity in Europe during the 19th century. A parlor game called Blotto asked participants to respond to inkblots, and a book by Kerner (1857) contained a collection of poetic responses to inkblot-like designs.

As the 19th century came to a close, several researchers began using inkblots in the study of various psychological operations. Krugman (1940) reported that Binet and Henri used inkblots to study visual imagination as early as 1895. Tulchin

(1940) noted that Dearborn's work at Harvard employed inkblots in the study of consciousness. Another American investigator, Whipple (1910) also used inkblots to study what he called "active imagination." A Russian researcher, Rybakow (1910) developed a series of eight inkblots to tap imagination. Hens, a staff member at Bleuler's clinic in Zurich, used inkblots with children, adult nonpatients, and psychiatric patients and did an unpublished dissertation on the topic.

The young Swiss psychiatrist, Hermann Rorschach, was thus not the first to include inkblots in the study of psychological processes when he began his project in 1911. But his work was qualitatively different from anything that had preceded it, in that he used inkblots to generate data from which extensive personality descriptions could be developed. Rorschach's preliminary but remarkably farsighted *Psychodiagnostik* was published in 1921. Tragically, he died within a year, at the age of 38, of complications of appendicitis.

It was three of Rorschach's friends, Walter Morgenthaler, Emil Oberholzer, and George Roemer, who insured that the insights and challenges of *Psychodiagnostik* were not lost. Morgenthaler had championed the book's publication in the face of some resistance from the Bircher publishing house. Oberholzer followed up by insuring that an important posthumous paper (Rorschach & Oberholzer, 1923) was published, and all three taught the test and encouraged adherents. One of Oberholzer's students, David Levy, took the test to Chicago, where he presented the first American Rorschach seminar in 1925.

Although each could have, neither Oberholzer nor Levy moved into a clear position as Rorschach's successor, and once in America the test was adopted by five psychologists of widely different backgrounds: Samuel Beck, Bruno Klopfer, Zygmunt Piotrowski, Marguerite Hertz, and David Rapaport. Of the five, only Beck, with the opportunity of a year's fellowship with Oberholzer in Zurich, was able to study with someone who had worked directly with Rorschach.

With little in the way of common background, the five Americans soon diverged in directions consistent with their theoretical orientations. They ultimately created five overlapping but independent Rorschach systems, each attracting adherents and each generating a body of literature and clinical lore. The history of the Rorschach from the late 1920s to the early 1970s is, to a large extent, the history of the development and elaboration of these five diverging systems.

Beck completed the first American Rorschach dissertation in 1932. He followed with a number of journal articles and published his *Introduction to the Rorschach Method* in 1937. He elaborated his system with additional books in 1944, 1945, and 1952, with updated editions published through 1967.

Klopfer had his first direct contact with the Rorschach in 1933. After a series of articles that included a description of a scoring system (Klopfer & Sender, 1936), he published *The Rorschach Technique* with Douglas Kelley in 1942. Elaborations of his system occurred in books published in 1954, 1956, and 1970.

Piotrowski was a member of a seminar offered by Klopfer in 1934, but within two years he had moved toward the creation of an independent system. His work culminated with the publication of *Perceptanalysis* in 1957.

Hertz, who worked briefly with Levy and Beck, included the Rorschach in her dissertation in the early 1930s, and continued research with the test for a decade after at the Brush Foundation in Cleveland. Sadly, the nearly 3000 cases she had amassed as well as an almost-completed manuscript describing her system were inadvertently destroyed when the Foundation closed. Although she never produced another book, her steady stream of journal articles and ongoing seminars led to the consolidation of the Hertz system by 1945.

Rapaport became interested in the Rorschach in the late 1930s and published a paper describing it in detail as part of a review of projective techniques in 1942. The first volume of *Diagnostic Psychological Testing* was published with Merton Gill and Roy Schafer in 1945, with the second volume following a year later. Schafer extended the system with additional books in 1948, 1954, and 1967, and Robert Holt edited a revised edition of the original two volumes in 1968.

With the publication of Piotrowski's book in 1957, all five of the systems were essentially complete. Each was taught independently and, during this period of divergence, each developed its own body of research and clinical literature. When Exner did a comprehensive review of the systems in 1969, he concluded that there were five overlapping but independent tests. Each of the systematizers had taken Rorschach's 10 inkblots and incorporated some of the ideas in *Psychodiagnostik* to create an instrument consistent with his or her theoretical stance. Each asked the person to respond to the cards and each attempted some sort of inquiry to clarify how the response was generated. Each had developed a format for coding various aspects of each response, its location and content, for example. And each system then generated an interpretation derived from the data that had been gathered. But, at every level from administration to interpretation there were major differences among the five systems.

These differences made sense when understood in the context of the varying theoretical positions and methodologies that the five systematizers brought to the Rorschach. At the coding level, Beck's rigorous positivism and behavioral training emerged in his insistence on normative and validation backing for the various elements. Klopfer's phenomenological background gave examiners greater leeway in using their own experience for reference. Rapaport, Hertz, and Piotrowski used methodological approaches between the extremes of Beck and Klopfer. At the interpretive level, Rapaport's extensive inclusion of psychoanalytic concepts separated his work from the less explicitly theory-based interpretive strategies of the other four systems.

Describing all the ways the five systems differ from one another is beyond the scope of this chapter. But a distinction suggested by [Weiner \(1977\)](#) identifies the critical question whose answer allows the characterization of an approach: How does the system conceptualize the nature of Rorschach data? Weiner suggested that the Rorschach can be viewed either as a perceptual–cognitive task or as a stimulus to fantasy.

The perceptual–cognitive stance assumes that the basic Rorschach task is to structure and organize a moderately ambiguous stimulus field. The way a person accomplishes this task is directly *representative* of real-world behavior that the

person will demonstrate in other situations requiring the same kind of operations. For example, people who solve the inkblots by breaking them into details which they combine into meaningful relationships (“two women stirring a pot of soup with a butterfly gliding by”) might be expected to engage day-to-day tasks in a similarly energetic, integrative manner.

The focus in the perceptual–cognitive conceptualization of the Rorschach is not on the words but rather on the structure of the responses, such as choice of location or integration of blot areas. This reliably quantifiable description of Rorschach response structure is utilized to generate descriptions of how the person is likely to behave in day-to-day life. These descriptions are based on the body of validity studies that link Rorschach structural variables with real-world behavior.

The stimulus-to-fantasy approach, on the other hand, views the Rorschach as an opportunity for the person to project material about internal states onto the ambiguity of the blots. The person’s responses are seen as *symbolic* of internal dynamics. As an example, a focus on the content of “two desperately disappointed people” might yield a dynamic description of internal conflict about interpersonal relationships.

The focus in the stimulus-to-fantasy approach is on the actual words, and this content helps create hypotheses not about likely behavior but rather about internal dynamics. The interpreter utilizes his or her theoretical framework and clinical experience to link Rorschach content and internal dynamics.

There is some question about where Rorschach himself should be placed in the perceptual–cognitive to stimulus-to-fantasy continuum. It is likely that he would have taken a middle-ground position that included both approaches. He criticized Hens’ 1917 work for its sole focus on content and imagination. In doing this, he differentiated himself from Hens and, by implication, from most of the earlier inkblot researchers, whose focus had been on verbalization and creative function. Rorschach stated that his primary interest was what he called “the pattern of perceptive process,” as opposed to the content of responses. *Psychodiagnostik* itself ([Rorschach, 1921/1942](#)) is almost entirely in the perceptual–cognitive camp, with particular attention to form, movement, and color. The 1923 posthumous paper added the structural element of shading, the light–dark contrast in some detail areas.

And yet Rorschach was trained in the work of Freud and Jung. He certainly would have been comfortable with the psychoanalytic description of projection as a mechanism through which individuals endow external material with aspects of their own dynamics, and with Frank’s classic 1939 paper that suggested that ambiguous stimuli such as inkblots could serve as “projective methods” for evoking this process. Indeed [Roemer \(1967\)](#) notes that Rorschach saw value in content analysis, citing a 1921 letter suggesting that he envisioned the technique as including both structural and symbolic material.

A review of their work suggests that all five of the American systematizers saw Rorschach data as having both perceptual–cognitive and stimulus-to-fantasy components, but with differences in relative emphasis. Beck stayed closest to the structural aspects. Rapaport was most willing to emphasize content, but even he wrote

“one can learn more about the subject sometimes by looking at a response from the point of view of its perceptual organization and at other times by looking at it from the point of view of the associative processes that brought it forth” ([Rapaport, Gill, & Schafer, 1946](#), p. 274).

Despite their overlap, each of the systems solidified and developed specialized terminology and literature. Clinicians schooled in one system found it increasingly difficult to communicate with those trained in other approaches as each system went its own way.

It was John Exner who, finding this level of diversity among the five systems in his 1969 review, undertook to provide the Rorschach community with a common methodology, terminology, and literature: the Comprehensive System (CS). In creating the CS, Exner reviewed the accumulated literature of all five systems and initiated some new research as well ([Exner, 1974, 1978, 1991, 1993, 2003, 2007; Exner & Weiner, 1986, 1995](#)). Using reliability and validity as criteria for inclusion, Exner’s project yielded a constellation of empirically defensible elements that forms the structural basis of the system. Content analysis is a secondary but important part of the CS. The approach to the handling of data as symbolic material can be characterized as dynamic but not specifically linked to any single theory of personality.

Exner first presented the Comprehensive System in 1974, and over the next decades the CS became the most widely used system throughout the American and international Rorschach communities. After his death in 2006, a group of Rorschach researchers ([Meyer, Viglione, Mihura, Erard, & Erdberg, 2011](#)) followed in his footsteps by creating a new system, the Rorschach Performance Assessment System (R-PAS), that incorporated the accumulated literature and undertook a variety of research projects to provide a contemporary infrastructure for the text. At this writing (2016) both the CS and R-PAS continue in use throughout the Rorschach community. What follows is a description of current research for both systems and an update of clinical developments.

## Current Rorschach research

### *The neuroscience of the Rorschach*

A number of studies have investigated the neurophysiological underpinning of various types of Rorschach responses, using neuroimaging or electroencephalographic technology.

Asari and his colleagues ([Asari et al., 2008, 2010](#)) at the University of Tokyo Medical School were interested in the neural substrate that underlies unique versus conventional perception. They had functional magnetic resonance (fMRI) technology available and used the Rorschach to evoke both unique and conventional percepts so that they could observe differential brain activation as a function of these different kinds of percepts. Administering the Rorschach while their participants were in the fMRI tunnel, the researchers used Japanese normative data as a control

against which to classify the 68 healthy participants' responses. They categorized responses as "unique," "infrequent," or "frequent." "Unique" responses did not occur in the normative control group, "infrequent" responses occurred in fewer than 2% of the control group's percepts, and "frequent" responses occurred in 2% or more of the control group's percepts. The researchers found significantly different fMRI activation patterns as a function of the level of uniqueness of Rorschach percepts. The study revealed clear-cut activation in the right temporal pole associated with unique perception. The right temporal pole is close to the limbic system with dense anatomic connections to the amygdala, often thought to be an important area for the convergence of emotional and perceptual signals. As noted, the right temporal pole activates with unique responses and actually closes down when the person gives either infrequent or frequent answers, with the differences significant at either the .05 or .01 levels. The anterior prefrontal regions, areas typically associated with executive function, activate above baseline with any response but activate strongly with conventional ('frequent') answers, significantly greater ( $P < .05$ ) than with infrequent answers. As responses move toward conventional, the prefrontal regions activate, and as they move toward uniqueness, the right temporal pole activates. The work of Asari and his colleagues suggests that reality-based versus fantasy-based percepts are generated in different brain areas, consistent with our understanding of the functions of these areas and confirmatory of the Rorschach's ability to document this critically important distinction.

Porcelli and his colleagues (Porcelli, Giromini, Parolin, Pineda, & Viglione, 2013) investigated the relationship between mirroring activity in the brain and Rorschach determinants. Mirror neurons are premotor cortical neurons that activate both during the execution of motor behavior and during its observation. Activity in the human mirror neuron system can be demonstrated by electroencephalogram (EEG) changes in the 8–13 µm frequency band. Porcelli et al. hypothesized that these mu changes would occur only with M responses, not with other Rorschach determinants or nonmoving human content, and would occur with greater intensity with accurate as opposed to less adequate M responses (less accurate form quality, passive movement, or nonwhole human content). They found that Rorschach responses involving human movement determinants (M) were uniquely associated with mirror neuron system activation, more strongly so if the action was active as opposed to passive. The authors concluded that the unique interpretation of Rorschach human movement responses as demonstrating empathy and social cognition is supported at the neurobiological level.

### **Rorschach variable selection**

An ongoing challenge for psychological test variables is the assessment of accumulating evidence for their validity as supported by the peer-reviewed literature. In 2012, Mihura, Meyer, Dumitrascu, and Bombel published a landmark meta-analysis in which they reviewed the 65 main variables in the Comprehensive System (CS). The authors tabulated 3074 validity coefficients, of which 1156 were judged to represent the variable's core construct. The effects came from 215 independent

samples, with a combined sample size of 25,795 participants. Most of the samples were from adult participants (152), but there were also 10 child, 31 adolescent, and 22 mixed age samples.

The authors found that variables with the strongest validity support were those that assessed distorted reality testing and serious cognitive slippage. These variables are important in their ability to identify and differentiate individuals with psychotic disorders. Variables with medium effect size relationships with validity criteria included some that assess available psychological resources and cognitive complexity (Lambda, Experience Actual, Human Movement, Difference Score, Complexity Ratio, Synthesized Responses, and Organizational Frequency). These variables are associated with ability to participate in psychotherapy and positive prediction of treatment outcome.

Other variables with positive validity support included the ratio of form dominance to color dominance, the Suicide Constellation, Cooperative Movement, Morbid content, and Anatomy and X-ray content.

It is also useful to review the CS variables that had the least validity support. Some 25 of the 65 CS variables fell into this category, as a function of (1) absence of validity studies, (2) nonsignificant validity findings, or (3) low or unstable validity findings. Many of these variables (e.g., color projection on achromatic cards) have very low base rates, leaving research on them virtually impossible.

Stressing the importance of multimethod approaches to clinical assessment, the authors suggest that valid Rorschach scores should contribute incremental validity over the MMPI.

The results from the Mihura et al. meta-analysis were the most significant source of data in deciding on variable inclusion or exclusion in the new Rorschach Performance Assessment System (R-PAS).

### ***Normative and form quality research***

An ongoing challenge for psychological tests is insuring that their normative data neither over- nor under-pathologizes respondents. A special supplement to the *Journal of Personality Assessment* (Shaffer, Erdberg, & Meyer, 2007) presented child, adolescent, and adult nonpatient Comprehensive System reference samples from around the world in order to provide contemporary norms. In a summary article to the supplement, Meyer, Erdberg, and Shaffer (2007) synthesized data from 21 adult samples from 17 countries in order to generate the Composite International Reference Values (CIRVs). A recent study (Meyer, Shaffer, Erdberg, & Horn, 2015) found that the CIRV norms are essentially identical when divided into three groups based on the quality of their data collection methodology. In a comparison of the CIRV norms with the existing Comprehensive System norms, the authors found that the CS norms over-pathologize healthy nonpatients across several aspects of psychological function. In a comparison with four different countries, the authors documented how two sets of within-country norms varied significantly, thus compromising the use of local norms instead of the CIRVs. The authors

recommend that the CIRV norms be used in clinical practice in that they provide a reference that is generalizable across settings, languages, and cultures. A subset of the CIRV norms is used in the new Rorschach Performance Assessment System (Meyer et al., 2011).

Another challenge for the Rorschach involves insuring that, when respondents are evaluated for reality testing, they are compared to a contemporary form quality reference. The authors of the R-PAS approach (Meyer et al., 2011) undertook a new project as a way of insuring that form quality was based on current international standards. In creating the new form quality table, the authors used fit and frequency as their guidelines. Fit is a measure of how well the response matches the features of the inkblot at the location where the object is perceived. Frequency is a measure of how often the object is reported throughout various international samples. In order to describe fit, 13,031 objects were rated by an average of 9.9 of 569 judges from Brazil, China, Finland, Israel, Italy, Japan, Portugal, Romania, Taiwan, Turkey, and the United States. The judges rated fit for the objects on a dimensional five-point scale ranging from “1 = No. I can’t see it at all. Clearly, it’s a distortion” through “2 = Not really. I don’t really see that. Overall, it does not match the blot area” through “3 = A little. If I work at it. I can sort of see that” through “4 = Yes. I can see that. It matches the blot pretty well” to “5 = Definitely. I think it looks exactly or almost exactly like that.” For the fit ratings, objects with mean ratings of 2.4 or less were considered to be minus in terms of form accuracy, objects with average accuracy ratings between 2.5 and 3.4 were considered to be unusual, and objects with an average rating of 3.5 or more were rated as ordinary by form accuracy criteria. Frequency was measured by the occurrence of objects on adult samples from Argentina, Brazil, Italy, Japan, and Spain. The least weight was given to objects that did not appear in any of the five samples, modest weight was given to objects that were seen by at least 1.5% of the participants in one sample, and the most weight was given to objects that were identified by at least 1.5% of the participants in two or more of the international samples. Ultimately, 39.9% of the objects in the new table have different form quality designations when compared to the CS form quality tables. However, initial validity research indicates that the CS and the R-PAS form quality tables are comparable in their ability to detect psychotic spectrum disorders.

## New clinical developments

Several new developments enhancing the use of the Rorschach in clinical practice have occurred recently. These include a method for optimizing the number of answers respondents give, a method for assessing the complexity of a respondent’s protocol, a composite for assessing ego impairment, an approach to quantifying orality and dependency, and an approach to describing how the respondent represents the interpersonal field. Each of these developments will be discussed.

### **R-Optimized method**

It has long been recognized (e.g., [Cronbach, 1949](#)) that the number of responses can impact the distribution of many other Rorschach variables, significantly compromising their interpretation. A series of studies ([Dean, Viglione, Perry, & Meyer, 2007](#)) concluded that the optimal range for responses was between 18 and 27. The authors of the Rorschach Performance Assessment System ([Meyer et al., 2011](#)) investigated a series of four alternate administration formats to develop an approach, the R-optimized method, that lowers the frequency of both short and long records, bringing more protocols into the optimal range. The R-optimized approach introduces the test by encouraging respondents to give “two, or maybe three” responses to each card. On any card in which the respondent gives only one answer, the examiner reminds the respondent with the “two, or maybe three” instruction. If the respondent gives four responses, the examiner takes the card back and again reminds the respondent with the “two, or maybe three” guideline. Findings indicated ([Reese, Viglione, & Giromini, 2014](#)) that the R-optimized approach yields fewer records outside of the optimal range while not differing significantly with values obtained from records administered with the Comprehensive System guidelines.

### **Complexity**

Factor analysis has suggested ([Meyer, 1992](#)) that the “first factor” on the Rorschach is a group of variables associated with the quantity and quality of the respondent’s output. The authors of the Rorschach Performance Assessment System developed a composite, Complexity, which is a good marker for this first factor. The Complexity composite is an aggregated combination of variables from various parts of the response process that show how much engagement and sophistication the person demonstrated in his or her response. It is calculated by giving points for (1) what locations the person used and whether he or she did any sort of integrating two or more detail areas or involved white space; (2) the number of content categories the person used; and (3) the number of determinants the person employed. The average correlation between Complexity and the R-PAS variables is .33, and Complexity thus becomes the variable with the greatest ability to make one person look different from another.

From an interpretive standpoint, Complexity is a good indicator of how much engagement the person demonstrated. It is a composite with both state and trait components, and the interpretive challenge is deciding whether a person’s Complexity score is more a function of traits like intelligence, creativity, attention to detail, openness to experience and coping resources—or whether it is a function of transient variables such as level of interest, motivation, engagement, and impression management. As an example, low Complexity could be a function of trait-level constriction, trauma, depression, or insecurity, or it could be a function of state-level guardedness or positive impression management. A standard score for Complexity of less than 85 or greater than 115 places the person more than a

standard deviation outside of the expected range. It may be useful to ask the web-based scoring program to perform a complexity adjustment, comparing the respondent only to individuals with a similar level of Complexity.

### **Ego Impairment Index**

Perry and Viglione (1991) began the development of a Rorschach index that assessed psychological disturbance with both adults and children. More recently, Viglione, Perry, Giromini, and Meyer (2011) developed an updated version of the Ego Impairment Index (EII-3) consistent with the variables and administration guidelines for the Rorschach Performance Assessment System. The authors contend that the Ego Impairment Index can best be understood as an index of ideational impairment that has implications for a range of ego functions. Ideational impairment is the underlying issue that mediates the challenges that people encounter in their day-to-day interaction with the environment. The Ego Impairment Index has four main threads, elevations on each of which have the potential for interfering with activities of daily living.

The four threads in the Ego Impairment Index include difficulties in reality testing, thought disturbance and cognitive slippage, interpersonal inaccuracy, and breakthrough of primitive material. Reality testing difficulties are assessed as the percentage of minus form quality answers (FQ-%) increases. Thought disturbance and cognitive slippage are assessed as the weighted sum of cognitive codes (WSumCog) such as fabulized combinations (WSumCog) increases. Problematical interpersonal function is assessed as the proportion of poor human representations to undistorted human representation (PHR/GPHR) increases, and breakthrough of critical content is assessed as the frequency of imagery such as blood, explosions, fire, and sex increases. Viglione et al. (2011) compared EII-3 values for psychotic and nonpsychotic patients with nonpatients in a focused contrast analysis that yielded highly significant differentiation of the three groups.

### **Oral Dependency Language**

The Rorschach Oral Dependency Language scale (previously known in the literature as ROD for Rorschach Oral Dependency, Bornstein, 1996; Bornstein, Hill, Robinson, & Calabrese, 1996; Masling, Rabie, & Blondheim, 1967) is one of the most robust Rorschach variables from a validity standpoint. It is coded only from the Response Phase of Rorschach administration, emphasizing its psychodynamic origins as an implicit measure of orality or dependency. The focus is on the respondent's spontaneous language, and responses such as a "tomato bug" or "the tongue of a shoe" would be coded. One point is given for any response that contains one or more of the Oral Dependency Language (ODL) categories. There are two types of responses: one involving oral content such as food and drinks or food sources and the other involving dependency content such as a person saying prayers. The percentage of responses that have an ODL code is used to create a variable (ODL%) that has implications for the respondent's implicit level of dependency, a finding of

which the respondent may not be aware or may be unwilling to describe. This often contrasts with dependency measures gained from self-report methodologies, which describe explicit, self-attributed need states.

### ***Mutuality of Autonomy***

The Mutuality of Autonomy (MOA) variable was developed by Urist in 1977 and has since been utilized in a variety of settings. Originally a seven-point scale, R-PAS has simplified it into two types of responses—healthy (MAH) and pathological (MAP). Healthy responses involve cooperative behavior between two equal participants (“two people working together to make a sculpture”), while pathological responses involve an imbalance of power which is used in a destructive manner (“a tiger killing a deer”). Based on object relations theory, the MOA provides information about how the person represents self, other, and the intervening relationship. This becomes a particularly important variable in the assessment of personality disorder, which frequently involves problematic representations of self and other and the relationship that characterizes the two.

A recent meta-analysis (Graceffo, Mihura, & Meyer, 2014) investigated 91 validity coefficients with a total of 1801 participants from 29 studies and 24 samples. The authors found a global effect size of .20 between MOA and any relevant criterion variable and some support for the R-PAS MAH and MAP distinction.

## **Conclusion**

As the Rorschach nears its 100th anniversary, perhaps its survival through many turbulent years can be traced to its ability to tap the complexity of human psychological operation. Neuroscience methodologies increasingly allow an understanding of the test’s neurophysiological infrastructure, supporting interpretation at a more and more nuanced level. The divergence that characterized the test for its first half century has been replaced by integrative approaches such as the Comprehensive System and the Rorschach Performance Assessment System. The years ahead should be productive ones for Rorschach’s deceptively simple technique.

## **References**

- Asari, T., Konishi, S., Jimura, K., Chikazoe, J., Nakamura, N., & Miyashita, Y. (2008). Right temporopolar activation associated with unique perception. *Neuroimage*, 41(1), 145–152.
- Asari, T., Konishi, S., Jimura, K., Chikazoe, J., Nakamura, N., & Miyashita, Y. (2010). Amygdalar enlargement associated with unique perception. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, 46(1), 94–99. Available from <https://doi.org/10.1016/j.cortex.2008.08.001>.

- Bornstein, R. F. (1996). Construct validity of the Rorschach Oral Dependency Scale: 1967–1995. *Psychological Assessment, 8*(2), 200–205.
- Bornstein, R. F., Hill, E. L., Robinson, K. J., Calabrese, C., et al. (1996). Internal reliability of Rorschach Oral Dependency Scale scores. *Educational and Psychological Measurement, 56*(1), 130–138.
- Cronbach, L. J. (1949). Statistical methods applied to Rorschach scores: A review. *Psychological Bulletin, 46*(5), 393–429.
- Dean, K. L., Viglione, D. J., Perry, W., & Meyer, G. J. (2007). A method to optimize the response range while maintaining Rorschach Comprehensive System validity. *Journal of Personality Assessment, 89*(2), 149–161.
- Exner, J. E. (1974). *The Rorschach: A comprehensive system*. New York: John Wiley & Sons.
- Exner, J. E., Jr. (1978). *The Rorschach: A comprehensive system: Vol. 2. Current research and advanced interpretation*. New York: John Wiley & Sons.
- Exner, J. E., Jr. (1991). *The Rorschach: A comprehensive system, Vol. 2. Interpretation* (2nd ed). New Tijrj: John Wiley & Sons.
- Exner, J. E., Jr. (1993). *The Rorschach: A comprehensive system, Vol. 1. Basic foundations* (3rd ed.). New York: John Wiley& Sons.
- Exner, J. E. (2003). *The Rorschach: A comprehensive system* (4th ed). New York: John Wiley & Sons.
- Exner, J. E., Jr. (2007). A new U.S. adult nonpatient sample. *Journal of Personality Assessment, 89*(Suppl. 1), S154–158. Available from <https://doi.org/10.1080/00223890701583523>.
- Exner, J. E., & Weiner, I. B. (1986). *The Rorschach : A comprehensive system* (2nd ed). New York: John Wiley & Sons.
- Exner, J. E., Jr., & Weiner, I. B. (1995). *The Rorschach: A comprehensive system: Vol. 3. Assessment of children and adolescents* (2nd ed). New York: John Wiley & Sons.
- Graceffo, R. A., Mihura, J. L., & Meyer, G. J. (2014). A meta-analysis of an implicit measure of personality functioning: The Mutuality of Autonomy Scale. *Journal of Personality Assessment, 96*(6), 581–595. Available from <https://doi.org/10.1080/00223891.2014.919299>.
- Kerner, J. (1857). *Klexographien*. Berlin: Boag and Co.
- Klopfer, B., & Sender, S. (1936). A system of refined scoring symbols. *Rorschach Research Exchange, 1*, 19–22.
- Krugman, M. (1940). Out of the ink-well: The Rorschach method [Also in Rorschach Res. Exch., 1940, 4, 91–101]. *Character & Personality: A Quarterly for Psychodiagnostic & Allied Studies, 9*, 91–110. Retrieved from <http://www.blackwellpublishing.com>.
- Masling, J., Rabie, L., & Blondheim, S. H. (1967). Obesity, level of aspiration, and Rorschach and TAT measures of oral dependence. *Journal of Consulting Psychology, 31*(3), 233–239.
- Meyer, G. J. (1992). The Rorschach's factor structure: A contemporary investigation and historical review. *Journal of Personality Assessment, 59*(1), 117–136.
- 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), S201–S216.
- Meyer, G. J., Shaffer, T. W., Erdberg, P., & Horn, S. L. (2015). Addressing issues in the development and use of the composite international reference values as Rorschach norms for adults. *Journal of Personality Assessment, 97*(4), 330–347. Available from <https://doi.org/10.1080/00223891.2014.961603>.

- 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.
- Perry, W., & Viglione, D. J. (1991). The Ego Impairment Index as a predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment*, 56(3), 487–501.
- Piotrowski, Z. A. (1957). *Perceptanalysys: A fundamentally reworked, expanded, and systematized Rorschach method*. Oxford, England: Macmillan.
- Porcelli, P., Giromini, L., Parolin, L., Pineda, J. A., & Viglione, D. J. (2013). Mirroring activity in the brain and movement determinant in the Rorschach test. *Journal of Personality Assessment*, 95(5), 444–456. Available from <https://doi.org/10.1080/00223891.2013.775136>.
- Rapaport, D., Gill, M., & Schafer, R. (1946). *Diagnostic psychological testing: The theory, statistical evaluation, and diagnostic application of a battery of tests: Vol. II*. Chicago: The Year Book Publishers.
- Reese, J. B., Viglione, D. J., & Giromini, L. (2014). A comparison between comprehensive system and an early version of the Rorschach performance assessment system administration with outpatient children and adolescents. *Journal of Personality Assessment*. Available from <https://doi.org/10.1080/00223891.2014.889700>.
- Roemer, G. A. (1967). The Rorschach and Roemer symbol test series. *Journal of Nervous and Mental Disorders*, 144(3), 185–197.
- Rorschach, H. (1921/1942). *Psychodiagnostics: A diagnostic test based on perception*. Oxford, England: Grune and Stratton.
- Rorschach, H., & Oberholzer, E. (1923). The application of the form interpretation test to psychoanalysis. *Zeitschrift für die Gesamte Neurologie und Psychiatrie*, 82, 240–274.
- Rybakov, T. (1910). *Atlas for experimental research on personality*. Moscow: University of Moscow.
- Shaffer, T. W., Erdberg, P., & Meyer, G. J. (2007). Introduction to the JPA special supplement on international reference samples for the rorschach comprehensive system. *Journal of Personality Assessment*, 89(Suppl. 1), S2–S6.
- Tulchin, S. H. (1940). The pre-Rorschach use of ink-blot tests. *Rorschach Research Exchange*, 4, 1–7.
- Viglione, D., Perry, W., Giromini, L., & Meyer, G. (2011). Revising the Rorshach ego impairment index to accommodate recent recommendations about improving Rorschach validity. *International Journal of Testing*, 11(4), 349–364.
- Whipple, G. M. (1910). *Manual of mental and physical tests*. Warwick and York: Baltimore.
- Zubin, J., Eron, L. D., & Schumer, F. (1965). *An experimental approach to projective techniques*. Oxford, England: Wiley.

## Further reading

- Beck, S. J. (1937). Introduction to the Rorschach method: A manual of personality study. *American Orthopsychiatric Association Monographs, No. 1*.
- Beck, S. J. (1944). *Rorschach's test: Vol. I. Basic processes*. Oxford, England: Grune & Stratton.
- Beck, S. J. (1945). *Rorschach's test: Vol. II. A variety of personality pictures*. Oxford, England: Grune & Stratton.

- Beck, S. J. (1952). *Rorschach's test: Vol. III. Advances in interpretation*. Oxford, England: Grune & Stratton.
- Beck, S. J. (1960). *The Rorschach experiment: Ventures in blind diagnosis*. Oxford, England: Grune & Stratton.
- Beck, S. J., Beck, A. G., Levitt, E. E., & Molish, H. B. (1961a). *Rorschach's test: I. basic processes* (3rd ed. revised). Oxford, England: Grune and Stratton.
- Beck, S. J., Beck, A. G., Levitt, E. E., & Molish, H. B. (1961b). *Rorschach's test: Vol. I. Basic processes* (3rd ed.). Oxford, England: Grune & Stratton.
- Exner, J. E. (1969). *The Rorschach systems*. New York: Grune & Stratton.
- Klopfer, B. (1970). *Developments in the Rorschach technique*. Fort Worth: Harcourt Brace.
- Klopfer, B., Ainsworth, M. D., Klopfer, W. G., & Holt, R. R. (1954). *Developments in the Rorschach technique: Vol. I. Technique and theory*. Yonkers-on-Hudson, NY: World Book Company.
- Klopfer, B. (1956). *Developments in the Rorschach technique: Vol. II. Fields of application*. Yonkers-on-Hudson, NY: World Book Company.
- Klopfer, B., & Kelley, D. M. (1942). *The Rorschach technique*. Oxford, England: World Book Company.
- Mihura, J. L., Meyer, G. J., Dumitrescu, N., & Bombel, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the comprehensive system. *Psychological Bulletin*, 139(3), 548–605. Available from <https://doi.org/10.1037/a0029406>.
- Rapaport, D. (1944). Manual of diagnostic psychological testing. 1: Diagnostic testing of intelligence and concept formation. *Josiah Macy Jr. Foundation Publications*.
- Rapaport, D., & Schafer, R. (1945). The Rorschach test: A clinical evaluation. *Bulletin of the Menninger Clinic*, 9, 73–77.
- Rapaport, D., Schafer, R., & Gill, M. (1946). Manual of diagnostic testing: II. Diagnostic testing of personality and ideational content. *Josiah Macy Jr. Foundation Publications Review Series*, 3.
- Urist, J. (1977). The Rorschach test and the assessment of object relations. *Journal of Personality Assessment*, 41(1), 3–9. Available from [https://doi.org/10.1207/s15327752jpa4101\\_1](https://doi.org/10.1207/s15327752jpa4101_1).
- Weiner, I. B. (1977). Approaches to Rorschach validation. In M. A. Rickers-Ovsiankina (Ed.), *Rorschach psychology*. Huntington, NY: Krieger.

## **Part VIII**

# **Behavioral Assessment**

# Behavioral assessment of children

15

Ross W. Greene and Thomas H. Ollendick

Department of Psychology, Virginia Tech, Blacksburg, VA, United States

Behavioral assessments of children can be characterized by several critical features. First, they are guided by a broad *social learning theory framework*. This framework, which is described more fully in the pages that follow, is distinguished by its empirical, data based approach to the study of persons. Consistent with this emphasis, behavioral assessments rely predominantly upon *objective* data and minimally upon subjective data or high levels of inference. Thus, a behavioral assessment differs from, for example, a dynamic formulation. While both might involve hypotheses regarding mechanisms underlying a child's patterns of thought and behavior, the latter is typically characterized by psychodynamic concepts for which confirmatory objective data are scarce. The emphasis on objectivity also necessitates consideration of *developmental* and *cultural norms*, and has ramifications for the selection of assessment procedures and for the types of conclusions one may draw from the information obtained through the assessment process. While "objective measurement" has, in the behavior analytic tradition, previously referred only to the direct observation of highly discrete overt behaviors and the identification of their controlling variables, assessments adhering to this orientation tend to be unnecessarily limited in scope and utility. From a social learning perspective, behavioral assessments of children focus upon both *overt* and *covert* behaviors, with the latter referring to affective states and various cognitive processes (e.g., expectancies, perceptions, beliefs, subjective values) that may exert considerable influence on overt behavior. As such, objective measurement refers to a broad array of assessment instruments and practices tapping overt and covert domains.

Second, behavioral assessments are distinguished by their *breadth* and *comprehensiveness*. This emphasis stems from an understanding that situational factors exert a powerful influence on the frequency, intensity, and duration of a given behavior. In this respect, a behavioral assessment differs from a diagnostic assessment, the primary goal of which is to determine the psychiatric categories that summarize the behaviors a child exhibits when he or she is having difficulty meeting expectations. Diagnostic assessments provide little highly-specific information about *contexts* contributing to variations in a child's behavior. Simply stating that a child "has attention deficit hyperactivity disorder (ADHD)" provides little or no information about the contexts in which the hyperactivity, impulsiveness, and/or

inattention that characterize this disorder are most problematic, nor any information about specific expectations the child is having difficulty meeting. Similarly, youth diagnosed with oppositional defiant disorder (ODD) are typically not oppositional most of the time; thus, specification of the disorder alone does not provide information about the specific contexts in which oppositional behavior occurs in individual children. The focus of a behavioral assessment is on what the child *does* in a given *situation* rather than on what he or she *has* or *is* (Mischel, 1968). Consequently, behavioral assessments extend well beyond the overt and covert behaviors of an identified child, and encompass the multiple persons (e.g., parents, teachers, siblings, peers) who interact (and have interacted previously) with the child, the multiple settings (e.g., various contexts within the home, school, and other environments) in which these interactions occur (or have occurred), and the expectations being placed upon the child by those persons and in those settings. As such, behavioral assessments are consistent with the themes of developmental psychopathology (e.g., Cicchetti, 1984, 1993; Rutter & Garmezy, 1983) and with theoretical models of development emphasizing reciprocal influences on behavior (e.g., Sameroff & Chandler, 1975).

The emphasis on breadth and comprehensiveness—combined with the limitations of different assessment procedures—also means that no single procedure or reporter is sufficient to provide a comprehensive assessment. Thus, behavioral assessments are *multimodal*, meaning that multiple reporters and measurement procedures are employed. In general, assessment conclusions are likely to be more definitive if there is stability in a child's behavior across time and contexts and general agreement across reporters and assessment procedures. In instances of inconsistency—for example, certain behaviors appear to be more frequent and intense in interactions within a particular context—the evaluation process must focus on factors—for example, different demands and expectations being placed on the child in different contexts—that may account for this inconsistency. Pinpointing these factors permits fine-tuned selection of intervention options.

Third, a behavioral assessment is best conceptualized as a *fluid, exploratory, hypothesis-testing process*, the goal of which is to *understand* a given child, group, family, and/or social ecology, often for the purpose of formulating and evaluating specific intervention strategies (McLeod, Jensen-Doss, & Ollendick, 2013; Ollendick & Greene, 1990; Ollendick & Hersen, 1984). In this respect, a behavioral assessment differs from, for example, a one-time mental status examination, in which the goal is to achieve an efficient summary of a child's mental state, often for the purpose of determining whether the child requires immediate psychiatric hospitalization. While mental status examinations clearly serve a critical function, the time-limited nature and urgency of such evaluations often force overly definitive, concrete, narrow conclusions driven by necessity of categorization (e.g., "this child is not psychotic and reports no suicidal or homicidal ideation or intent; while his behavior is of concern and may signal an emerging bipolar illness, he does not require hospitalization at this time"). The time constraints involved in a mental status examination may also encourage the invocation of fairly amorphous explanations and recommendations (e.g., "clearly the stress caused by the acrimonious

relationship between this child's mother and father is impacting upon his already tenuous emotional state; he and his parents have been referred to a mental health clinic in their community for consultation”).

Inherent in a behavioral assessment is the understanding that a definitive understanding of a child's overt and covert behavior is difficult to achieve. As such, initial conclusions are understood as *hypotheses* that await verification or revision based on additional information. The gathering of additional information is understood to be an ongoing, never-ending process. This process continues even during intervention, which, aside from being an attempt to ameliorate a child's difficulties, can be understood as an opportunity to obtain additional information based on the child's response to treatment. Thus, child behavioral assessments de-emphasize quick, definitive conclusions and focus more on obtaining information that is directly relevant to treatment. “Relevance for treatment” refers to the clinical utility of information in pinpointing treatment goals, the selection of targets for intervention, the design and implementation of interventions, and the evaluation of intervention outcomes (Greene, 1995; Mash & Terdal, 1988). A related concept—treatment utility—refers to the degree to which assessment strategies are shown to contribute to beneficial treatment outcomes (Hayes, Nelson, & Jarrett, 1987).

In sum, the behavioral assessment of children can be understood as *a fluid, exploratory, hypothesis-testing process guided by social learning principles in which a range of specific procedures is used in order to achieve a broad, comprehensive understanding of a given child, group, and social ecology, and to formulate and evaluate specific intervention strategies*. Before expanding upon the various assessment components delineated above, a brief historical overview may be useful.

## History of behavioral assessment

The history and evolution of behavioral assessment closely parallels the history and evolution of behaviorism. Behaviorism evolved at least in part as a reaction to psychoanalytic theory, which emphasized concepts viewed by behaviorists as subjective, unobservable and, therefore, unscientific. One of the goals of the early behaviorists was to elucidate a philosophy of understanding and studying persons driven by the objective procedures that typified the “harder” sciences. From a behavioral perspective, psychoanalytic concepts such as drives, conflicts, psychosexual stages of development, unconscious defense mechanisms, and the like, could not be measured objectively and had to be *inferred* from a person's self-report or behavior. Thus, such concepts had no place in a scientific theory of human behavior and could not be invoked to explain mechanisms controlling a given behavior. Early behaviorists believed that only directly observable behaviors were worthy of study (Skinner, 1953). A person's self-report was viewed not only as unreliable but also as uninterpretable, since interpretation—including the inferences of the clinician—was viewed as the epitome of subjectivity. Direct observation of behavior was the hallmark of behavioral assessment. Thus, rather than interpreting

behaviors—for example, characterizing an oppositional child as “hostile” or “enraged”—behaviorists would instead strive to operationalize or define oppositional behavior in objective terms (e.g., “tantrums,” “screaming,” “swearing,” “refusing to comply with adult directives,” etc.) and embark on an assessment process that involved direct measurement of the frequency, intensity, and duration of these behaviors under varying conditions or situations.

By emphasizing science, behaviorists also stimulated much research examining the normative developmental course of children’s behavior. Whereas various interpretations had previously been applied to certain childhood behaviors (e.g., hitting is a sign of internal rage, bedwetting is a primitive attempt to extinguish this internal rage, clingy behavior is indicative of enmeshment, and so forth), researchers began to demonstrate that childhood behaviors once considered “deviant” were actually fairly normative at certain ages (e.g., it is not developmentally unusual for 2 year olds to be impulsive and inflexible, for 3 year olds to wet their bed at night, for 5 year olds to reverse letters, and for 12 year olds to be anxious about their physical appearance). Thus, the increased emphasis upon determining whether a particular behavior or pattern of behavior was *developmentally deviant* was by no means coincidental (see [Campbell, 1989](#), for further discussion of this issue).

Behaviorists also refuted notions of “personality” as consisting of stable and enduring traits and of behavior as consistent across situations and over time (we place “personality” in quotations because early behaviorists would have objected to use of this term, given its ambiguous parameters). As noted above, the behavioral view has instead emphasized the *situational* nature of behavior. In the case of a child who exhibits oppositional behavior, for example, a behaviorist would assume that such behavior does not occur at all times but rather under certain conditions and settings, and would therefore attempt to identify those variables that elicit and maintain such behavior in those situations. In other words, the assessment process would extend well beyond the clinic and would document the child’s behavior at various points in time at home, school, the playground, at hockey practice, and so forth. Recall that in a behavioral assessment the goal is not to make sweeping generalizations about a child’s behavior but rather to reach a clear understanding of the conditions under which certain specific behaviors occur.

Not coincidentally, these emphases on directly observable stimuli and situational factors led to the conclusion that behavior occurs by virtue of *conditioning* (i.e., *learning, experience*). By manipulating environmental conditions, behaviorists were able to demonstrate convincingly the manner in which behavior could be shaped, elicited, maintained, and eliminated. Briefly, in *classical conditioning* a neutral stimulus (one that elicits no particular response) is repeatedly presented along with a stimulus that reflexively elicits a particular response, with the neutral stimulus alone eventually eliciting the same response. In *operant conditioning* behavior is governed by its consequences; behaviors that are rewarded are more likely to be repeated whereas behaviors that are punished are less likely to be repeated.

While literally thousands of published scientific studies have offered compelling evidence for the role of classical and operant conditioning as influences on the behavior of humans and other animals, behaviorists came to recognize that other

factors influenced human behavior, driven by the influential work of Rotter (1954, 1966, 1972), Bandura (1971, 1973, 1986), Mischel (1968, 1973, 1979), and others. These *cognitive behaviorists* or *social learning theorists* extended the work of early behaviorists in a variety of ways, perhaps most significantly by introducing the notion that cognitions and affective states exert significant influence upon learning and behavior, and by proposing that much learning occurs indirectly by observing others (vicarious learning) rather than directly through classical and operant conditioning.

True to their scientific roots, social learning theorists proposed and attempted to study specific categories of cognition and the manner in which these cognitions influence learning and human behavior. Mischel (1973, 1984) referred to these categories as *cognitive social learning person variables* (CSLPVs) and, because of their significance for behavioral assessment, they are worthy of brief overview here. *Competencies* refers to a person's actual skills and capacities, be it intelligence, reading skills, social skills, problem solving skills, and so on; *encoding strategies* refer to the manner in which a person interprets and perceives themselves, others, and events; *expectancies* are a person's expectations of the likely outcomes of a particular behavior or stimulus and the degree to which a person believes he or she can actually perform a particular behavior; *subjective values* refers to a person's preferences and aversions, likes and dislikes, and so on (i.e., what stimuli are rewarding to the person and what stimuli are punishing); and *self-regulatory systems and plans* refers to a person's capacity for and manner of self-imposing goals and standards and self-administering consequences.

How might these social learning person variables be of value in the assessment of a child who exhibits oppositional behavior? Were we to assess the child's competencies, we might, for example, consider the child's capacities or skills for formulating behavioral alternatives (besides screaming) for dealing with the frustration that occurs in response to adult demands for compliance and modulating the emotions associated with frustration. The child's encoding strategies might also be relevant: Does the child believe that he or she is being treated unfairly in comparison to siblings or that he or she is being blamed for sibling interactions that are transactional rather than unidirectional? Also relevant are the child's expectancies: Does the child believe that compliance with adult expectations will lead to advantageous outcomes? Does the child believe he or she is capable of meeting these expectations? Are the expectations realistic? Information about CSLPVs should lead to a more comprehensive conceptualization of the factors driving a child's behavior, should identify situations in which behavior is more or less likely to occur, and might result in more productive interventions than, for example, merely punishing oppositional behavior and rewarding compliance with adult directives. As discussed below, consideration of CSLPVs may also lead to an improved understanding of the persons with whom a given child interacts.

The above discussion has significant implications for notions regarding the "function" of a child's maladaptive behavior. The traditional conceptualization of function is that challenging behavior is "working" at helping a child get something

(e.g., attention) and/or escape and avoid expectations that are challenging, tedious, unpleasant, or scary. An alternative view—one that incorporates a broader social learning perspective—is that challenging behavior is instead the means by which a child is *communicating* that he or she is having difficulty meeting certain demands and expectations. The latter definition of function would be more likely to facilitate consideration of CSLPVs and identification of the specific expectations the child is having difficulty meeting.

As we have already noted, social learning theory also proposes that much learning occurs through observational or vicarious processes rather than solely through classical and operant conditioning. In other words, tantrums may have entered a child's repertoire by observing others (e.g., at home, in the classroom, on television) rather than solely by the child having been reinforced for noncompliance. We may also learn by hearing or reading about others' behavior or outcomes. Thus, there are virtually limitless opportunities and situations through which behaviors may enter a child's repertoire, and all children—those who live in circumstances viewed as ideal and those living in circumstances viewed as less ideal—have advantageous and disadvantageous behaviors in their repertoires. Therefore, behavioral assessment of a child must extend into important environmental domains in which the child has had the opportunity to observe, hear about, and learn such behaviors. This may include large-scale social systems such as schools and neighborhoods (Patterson, 1976; Wahler, 1976) which have been shown to have immediate and profound effects on individual behavior (see Winett, Riley, King, & Altman, 1989, for a review of this issue). Although inclusion of these additional factors may add complexity to the assessment process, they are indispensable to the goal of achieving a broad, comprehensive, and clinically useful assessment of a child.

In sum, child behavioral assessment has evolved from sole reliance on measurement of directly observable target behaviors into a broader approach that takes into account cognitive and affective processes, developmental issues, and social contexts that contribute to variations in a child's behavior. Needless to say, multimethod, multicontext behavioral assessment of children entails use of a wide range of specific procedures. Let us now turn to an overview of how these procedures connect with the various components of a behavioral assessment reviewed in these introductory sections. The overview is not organized sequentially but rather by assessment domains: *overt behavior*, *covert behavior*, and *contexts*. Procedures used in the assessment of a child's overt behavior tend to focus primarily on the issue of "What does the child do, and when?" while procedures utilized in assessment of covert behavior focus on issues of "What does the child think and feel, and when?" and "How do these thoughts and feelings connect with the child's overt behavior?" Finally, procedures used in the assessment of contexts center on issues of "How do overt and covert behaviors and expectations of adult caretakers and characteristics of environments contribute to variations in a child's overt and covert behavior?" Given space limitations, we have chosen not to provide an exhaustive overview of assessment instruments; instead, we provide examples of various measures and procedures that may assist in achieving each component.

## Assessment of children's overt behavior

Various procedures may be employed in attempting to gather information about what overt behaviors a child exhibits and the conditions under which these behaviors occur. Early on, [Cone \(1978\)](#) distinguished between direct and indirect methods of assessing a child's overt behavior; in *direct* assessment, the behaviors of interest are assessed *at the time and place of their occurrence*, whereas *interviews* with the child or other reporters, *self-reports* by the child, and *behavior ratings* by the child or other reporters are considered more *indirect* means of obtaining assessment information. We discuss each of these methods and their relative advantages and limitations below.

### ***Direct behavioral observation***

Direct observation involves the formal or informal observation of a child's overt behavior in various natural contexts, including home (e.g., during homework, dinner, bedtime, etc.), school (e.g., during recess, lunch, group discussions, independent work, etc.), and other domains (little league, friends' homes, etc.). In addition to providing a first-hand view of behaviors of interest, such observations also afford an opportunity to observe situational factors contributing to variations in a child's behavior. [Johnson and Bolstad \(1973\)](#) have characterized the development of naturalistic observation procedures as the major contribution of the behavioral approach to assessment and treatment of children. While this point is arguable, it is clear that direct observation may involve the least inference of the assessment methods we will describe and remains an indispensable component of a behavioral assessment. That said, direct observations tend to be more time-consuming and inconvenient as compared to other methods of assessment, and there is no guarantee that target behaviors will actually occur during designated observation periods. Thus, in some cases it may be necessary to have significant others in the child's environment (e.g., parents, teachers) formally observe and record the child's behavior, to have children observe and record their own behavior, or to employ analog procedures in which the goal is to observe the child in a simulated laboratory setting that closely resembles the actual setting(s) of interest.

In many instances, it may be desirable to formalize the direct observation process, especially if the observer wishes to quantify behaviors for the purpose of normative comparisons. In such instances, a measure such as the Child Behavior Checklist Direct Observation Form (CBC-DOF; [Achenbach, 1986](#)) may be useful. This measure consists of 96 items covering a broad range of children's behavior. The form is completed after observing a child in a given setting for 10 min after a recommended three to six separate occasions. Scores can be obtained on six factor-analytically-derived scales, including withdrawn-inattentive, hyperactive, nervous-obsessive, depressed, attention demanding, and aggressive. An advantage of the CBC-DOF is its sensitivity to developmental issues (normative information based

on a child's age and gender is available) and sound psychometric properties ([McConaughy, Achenbach, & Gent, 1988](#)).

In other cases, observers may wish to quantify more specific realms of behavior. An example of a system designed for such purposes is the Classroom Observation Code developed by [Abikoff, Gittelman-Klein, and Klein \(1977\)](#) and modified by [Abikoff and Gittelman \(1985\)](#), which is used to record behaviors associated with poor self-regulation in school settings. The system involves the systematic recording of various categories of behavior reflective of poor self-regulation, including (but not limited to) *interference* (verbal or physical behaviors that are disturbing to others), *off-task* (attending to stimuli other than the assigned work), *noncompliance* (failure to follow teacher instructions), *minor motor movement* (e.g., restlessness and fidgeting), *gross motor behavior* (e.g., leaving seat and/or engaging in vigorous motor activity), *physical aggression* (physical aggression directed at another person or destruction of others' property), and *solicitation of teacher* (e.g., raising hand, calling out to teacher). Another example is the Revised Edition of the School Observation Coding System (REDSOCS; [Jacobs et al., 2000](#)), which classifies behaviors on three behavioral domains: (1) appropriate versus inappropriate behaviors, (2) compliant versus noncompliant behavior, and (3) on-task versus off-task behavior.

In yet other cases, assessors may wish to *simultaneously* record the behavior of children and the adults with whom they are interacting for the purpose of capturing the reciprocal nature of adult/child interactions. Because such recording systems tend to be labor intensive, their use is often restricted to research. However, one cannot overstate the conceptual appeal of conducting observations that attend simultaneously to child and adult behavior so as to obtain information about the manner in which the behavior of each may lead to variations in the behavior of the other, even if such observations occur informally and in unstructured contexts. For example, the Main-Stream Code for Instructional Structure and Student Academic Response (MS-CISSAR; [Greenwood, Carta, Kamps, Terry, & Delquadri, 1994](#)) is used to record student-teacher interactions and students' response to intervention. Instruments used to record interactions between children and parents include the Response Class Matrix ([Barkley, 1981](#); [Mash & Barkley, 1986](#)), the Dyadic Parent–Child Interaction Coding System IV (DPICS; [Eyberg, Nelson, Ginn, Bhuiyan, & Boggs, 2013](#)), and the Living in Family Environments system (LIFE; [Hops et al., 1990](#)).

As noted above, direct observation can also be accomplished via laboratory or analog settings that are similar to, but removed from, the natural environment. Simulated observations are especially helpful when a target behavior is of low frequency, when the target behavior is not observed in the naturalistic setting due to reactivity effects (i.e., the child does not exhibit the target behavior because he or she is aware of being observed), or when the target behavior is difficult to observe in the natural environment due to practical constraints. An example of such a simulated observation system is the Restricted Academic Situation (RAS) described by [Barkley \(1990\)](#), which has been used in the analog assessment of poorly self-regulated children. In brief, the RAS involves placing a child in a playroom

containing toys and a small worktable; the child is instructed to complete a packet of mathematics problems and is observed for 15–20 min from behind a two way mirror. Behavior coding occurs during this period, using clearly defined categories such as off-task, fidgeting, vocalizing, playing with objects, and out of seat. A disadvantage of this system is its lack of normative data, and the degree to which behaviors observed during this and other analog procedures generalize to natural settings is an important concern.

### ***Behavior ratings and checklists***

Behavior checklists are perhaps the most popular indirect means of gathering information about a child's overt behavior, and this popularity is presumably a function of the efficiency of such checklists. In their most common form, behavior checklists require adults to rate various items to indicate the degree to which a child exhibits certain behaviors. In general, checklists are useful in providing an overall description of a child's behavior, in specifying dimensions or response clusters that characterize the child's behavior, and in serving as outcome measures for the effectiveness of treatment. Again, this method of assessment is considered *indirect* because it relies on retrospective descriptions and ratings of the child's behavior by reporters whose impartiality is uncertain. Thus, as with direct observation, assessors must be sensitive to potential biases of persons completing and interpreting checklists. Nonetheless, behavior checklists can provide a comprehensive and cost-effective picture of the child and may be useful in eliciting information that may have been missed by other assessment procedures (Novick, Rosenfeld, Bloch, & Dawson, 1966). Many checklists undergo rigorous evaluations of reliability and validity and provide developmental norms. While normative information helps protect against some forms of subjectivity, behavior checklists should not be viewed as a satisfactory replacement for direct observation (though they are often used this way); rather the two methods of assessment are best viewed as complementary.

Among the most widely used of the "omnibus" checklists—those assessing a broad range of behaviors—are the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2000, 2001), which is completed by parents, and the Child Behavior Checklist Teacher Report Form (CBC-TRF; Achenbach & Rescorla, 2001), which is completed by teachers. The CBCL is available in two formats depending on the age of the child being rated (i.e., 1½–5 years and 6–18 years; we review only the latter here). The CBCL/6-18 consists of 112 items rated on a 3-point scale. Scored items can be clustered into three-factor analyzed profiles: social competence, adaptive functioning, and syndrome scales. The latter includes scales such as withdrawn, anxious/depressed, social problems, attention problems, delinquent behavior, and aggressive behavior. Social competency items examine the child's participation in various activities (e.g., sports, hobbies, chores) and social organizations (e.g., clubs, groups), and school (e.g., grades, placement, promotions). The teacher-completed CBC-TRF also consists of 112 items which are fairly similar to those found in the CBCL. The scored items from the CBC-TRF cluster into the same three-factor analyzed profiles. Both checklists have been extensively researched, provide detailed

normative information, and have exceptional psychometric properties. Some of the items on the CBC-TRF and CBCL refer to directly observable behaviors (e.g., physically attacks people, bites fingernails, gets teased a lot) whereas others refer to cognitions and affective states (e.g., fears he or she might think or do something bad, likes to be alone, feels too guilty, feels worthless or inferior). As such, while we have placed our discussion of these checklists in this section on assessment of *overt* behavior, we wish to emphasize that they also provide information about select *covert* processes.

Again, one should not assume that the excellent psychometric properties of the CBCL and CBC-TRF protect these instruments from the idiosyncratic perceptions and biases of raters. Many items are not clearly defined and do, in fact, require subjective judgments (e.g., acts too young for his or her age, demands a lot of attention, showing off or clowning, too fearful or anxious) and therefore may contribute to variability in ratings by different adults (see [Achenbach, McConaughy, & Howell, 1987](#)). Thus, variability in responses of different raters may reflect not only the influence of contextual factors (i.e., the child's behavior varies depending on the expectations or other characteristics of different situations) but also the unique interpretations, expectations, and tolerances of different adult raters ([Greene, 1995, 1996](#)). Unlike direct naturalistic observation, behavior checklists are one or more steps removed from the behaviors themselves, and it may be difficult to clarify responses and explore potential response biases of raters.

It is often useful to obtain more detailed information about a specific domain of behavior—such as social skills or self-regulation—than that provided by omnibus behavior checklists. Many “narrow band” checklists have been developed and may be useful in this regard. For example, the Social Skills Rating System (SSRS; [Gresham & Elliott, 1990](#)) is a 55-item questionnaire that provides information about a child's social behavior in three domains (social skills, problem behaviors, and academic competence). The SSRS includes parent-, teacher-, and self-rated formats, and the instrument has very good psychometric properties ([Diperna & Volpe, 2005](#)). Numerous checklists are available for obtaining ratings of a child's self-regulation, including the Behavior Rating of Executive Function (BRIEF; [Gioia, Isquith, Guy, & Kenworthy, 2000](#)), which has been recently updated (BRIEF-2) and includes parent-, teacher-, and self-report screening forms. This instrument has been extensively researched and also has excellent psychometric properties.

## **Interviews**

Of the many assessment procedures employed by behavioral clinicians, the interview is the most widely used ([Swann & MacDonald, 1978](#)) and is also considered an indispensable part of assessment ([Gross, 1984; Linehan, 1977; McConaughy, 1996](#)); so indispensable, in fact, that we discuss behavioral interviews not only in this section (as related to assessment of a child's overt behavior) but also in subsequent sections (related to assessment of a child's covert behavior and of the behavior and cognitions of others in a child's environments). Similar to naturalistic observations, such interviews may be *informal* or *formal*.

As an assessment of overt behavior, informal interviews may yield a wide range of detailed information about a child's specific behavior and much preliminary information about possible controlling variables. Such information may assist in the early formulation of treatment plans and in the development of relationships with the child and his or her family (Ollendick & Cerny, 1981). As with direct observation and behavior checklists, clinicians must recognize that the impartiality and objectivity of informants is uncertain. At the risk of redundancy, reporters—adults and children—are active interpreters of behavior; thus, the accuracy and reliability of information they provide always involves some level of subjectivity.

The popularity of interviews may derive in part from a number of practical considerations, as well as to advantages they offer over other procedures (Gross, 1984). As we have noted, direct observations of target behaviors are essential but frequently inconvenient and impractical. Moreover, behavior checklists often do not permit detailed exploration or elaboration of responses. An informal interview permits the clinician to efficiently obtain a broad band of information about overall functioning as well as detailed information about specific areas, and this information can be used as the initial basis for early hypotheses and interventions. Informal interviews also involve greater flexibility than formal interviewing methods described below, allowing the clinician to build a relationship with the child and his or her family and to obtain information that might otherwise not be revealed. As noted by Linehan (1977), some family members may be more likely to divulge information verbally in the context of a professional relationship than to write it down on a form to be entered into a permanent file.

Formal interviews may be either structured or semistructured. In general, structured interviews are oriented toward specific diagnostic categories and require parents (and/or children and adolescents) to endorse diagnostic items for the wide range of childhood disorders in the DSM-5 (American Psychiatric Association, 2013) or ICD-10-CM (World Health Organization, 2013). Clearly, such interviews facilitate collection of systematic data relative to a broad range of symptoms and diagnoses. However, structured interviews may be limited by (1) an overemphasis on diagnostic categories (to the exclusion of other important information), (2) weak self-report reliability for children under age 13, (3) low reliability between responses of children and parents, and (4) dichotomous present-or-absent scoring (McConaughy, 1996). Further, structured interviews typically do not yield specific information about contextual factors; thus, even when a structured interview is used, subsequent informal interviewing is often required for the purpose of clarifying responses. Several reliable, valid "omnibus" structured diagnostic interviews are available, such as the Diagnostic Interview Schedule for Children-IV (DISC-IV; Fisher, Lucas, Sarsfield, & Shaffer, 2006). Other structured interviews are oriented toward a specific domain, such as anxiety, as with the Anxiety Disorders Interview Schedule for Children (ADIS-IV; Silverman & Albano, 1996). Various semistructured interviews have been developed as an alternative to the rigid format and focus on categorical diagnoses that characterize structured interviews, including the Semistructured Clinical Interview for Children and Adolescents (SCICA; McConaughy & Achenbach, 2001), a child and adolescent self-report interview.

## Assessment of a child's covert processes

If one acknowledges the influence of cognitive and affective processes on overt behavior, then the importance of assessing these processes is self-evident. In addition to obtaining information from adult caretakers about a child's covert processes via interviews or behavior checklists, more direct methods for obtaining such information may also be useful. As described above CSLPVs—competencies, encoding strategies, expectancies, subjective values, and self-regulatory systems and plans—provide a framework for guiding the assessment of covert processes.

### **Interviews**

Clearly, directly interviewing a child may yield much information about covert processes, assuming the child is willing and able to report on such processes. As we have noted, early behaviorists eschewed such interviews, maintaining that observable behavior was the least inferential method of obtaining assessment information. To a large extent, such negative bias against self-report was an outgrowth of early findings indicating that reports of subjective states did not always coincide with observable behaviors (Finch & Rogers, 1984). While congruence in responding is, in fact, not always observed, contemporary researchers have cogently argued that a child's *perceptions* of his or her behavior and its consequences may be as important for behavior change as the behavior itself (Finch, Nelson, & Moss, 1983; Ollendick & Hersen, 1984). We are aware of no structured interviews used specifically for the purpose of assessing CSLPVs, but informal interviews may yield critical information about competencies ("Besides hitting, what else could you do when your brother enters your room without permission?"), encoding strategies ("Do the kids at school seem to like you?"), expectancies ("What would happen if you tried to join in on the four square game at school?"), subjective values ("How important is it to you to do well on that science test?"), and self-regulatory systems and plans ("Do you have trouble sitting still in class? Do you call out answers without raising your hand in class?").

### **Direct measures**

An excellent example of a broad-based instrument used in the assessment of competencies is the IQ test. While one could reasonably argue that an IQ test such as the Wechsler Intelligence Scale for Children—Fifth Edition (WISC-V; Wechsler, 2014) is, in fact, a sampling of overt behavior, there seems little question that the broad dimensions tapped by IQ measures go well beyond the single sampling of behavior obtained during testing. Indeed, IQ tests may provide significant information about a child's competencies in a wide range of areas, such as acquired knowledge, abstract logical reasoning skills, problem-solving skills, long- and short-term auditory and visual memory, social judgment, sequencing skills, and mental flexibility, each of which may impact quite directly on a child's behavior. The

impressive normative data available for measures such as the WISC-V add to its value and credibility as a direct measure of a broad range of competencies. Another instrument—the NEPSY-II (Korkman, Kirk, & Kemp, 2007a, 2007b)—is a comprehensive instrument designed to assess neuropsychological development in realms not typically covered by general ability or achievement batteries. It includes six functional domains, including attention and executive functioning, language, memory and learning, sensorimotor skills, social perception, and visuospatial Processing. The Clinical Evaluation of Language Fundamentals—Fifth Edition (CELF-5; Wiig, Semel, & Secord, 2013) has a somewhat narrower focus on language strengths and weaknesses, including oral and written language and nonverbal communication skills.

### ***Self-report instruments***

In addition to simply asking a child about covert processes in an interview format, various self-report instruments have also been developed to access this information. As with parent- and teacher-completed checklists, self-reports should be used with appropriate caution and due regard for their specific limitations. Because they generally involve the child's retrospective rating of attitudes, feelings, and behaviors, self-report instruments should be considered indirect methods of assessment (Cone, 1978).

Some self-report checklists focus on a broad range of overt and covert processes, as in the case of the Youth Self-Report (YSR; Achenbach & Rescorla, 2001). The YSR includes items very similar to the CBCL and CBC-TRF described earlier, and is appropriate for children and adolescents aged 11–18 years. While many YSR items are reflective of overt behaviors (e.g., I bite my fingernails; I destroy my own things; I have nightmares), it is perhaps most useful for assessing a variety of covert behaviors (e.g., I'm too dependent on adults; I don't feel guilty after doing something I shouldn't; I am afraid of going to school; I feel that I have to be perfect; I feel that others are out to get me; my moods or feelings change suddenly). As such, the YSR may be used to assess various of the CSLPVs (e.g., encoding strategies, subjective values, expectancies).

Other self-report checklists may tap a more specific area of interest, such as anxiety, which is comprised of both overt *and* covert behaviors. Indeed, anxiety may tap into numerous CSLPVs, including expectancies and encoding strategies. Moreover, some anxieties may not be manifested overtly in some children, and assessment methods which are primarily oriented toward overt behavior may therefore be less informative. Self-report scales such as the Revised Children's Manifest Anxiety Scale: Second Edition (RCMAS-2; Reynolds & Richmond, 2008), the State-Trait Anxiety Inventory for Children (STAIC; Spielberger, 1973), the Child Anxiety Sensitivity Index (CASI; Silverman, Fleisig, Rabian, & Peter son, 1991), the Social Anxiety Scale for Children-Revised (SASC-R; LaGreca & Stone, 1993), and the Fear Survey Schedule for Children-Revised (FSSR-R; Ollendick, 1983; Ollendick, Matson, & Helsel, 1985) may be extremely useful for assessing covert aspects of children's anxieties.

For example, the FSSC-R (Ollendick, 1983) is the revised version of the Fear Survey Schedule for Children (Scherer & Nakamura, 1968). In the revised scale, designed to be used with younger and mid-age children, the child is instructed to rate his or her fear level to each of 80 items on a 3-point scale. Children are asked to indicate whether a specific fear item (e.g., having to go to school, being punished by father, dark places, riding in a car) frightens them “not at all,” “some,” or “a lot.” Factor analysis of the scale has revealed five primary factors: fear of failure or criticism, fear of the unknown, fear of injury and small animals, fear of danger and death, and medical fears (Ollendick, King, & Frary, 1989). Moreover, it has been shown that girls report greater fear than boys, that specific fears change developmentally, and that the most prevalent fears of boys and girls have remained unchanged over the past 30 years. Further, studies have shown that the instrument is sensitive to cultural influences (Ollendick, Yang, King, Dong, & Akande, 1996). The instrument can be used to differentiate subtypes of specifically phobic youngsters whose fear of school is related to separation anxiety (e.g., death, having parents argue, being alone) from those whose fear is due to specific aspects of the school situation (e.g., taking a test, making a mistake, being sent to the principal). Recently, a 25-item short form of this instrument has been developed which shows similar psychometric properties (Muris, Ollendick, Roelofs, & Austin, 2014).

Still other instruments are available to tap a broad range of additional covert processes, such as self-concept (e.g., the Piers–Harris Children’s Self-Concept Scale; Piers, 1984), self-perception (the Self-Perception Profile for Children; Harter, 1985, and the Self-Perception Profile for Adolescents; Harter, 1988); depression (e.g., the Children’s Depression Inventory-2; Kovacs, 2014), and locus-of-control (e.g., the Nowicki–Strickland Locus of Control Scale for Children; Nowicki & Strickland, 1973).

### ***Self-monitoring***

Self-monitoring can be used to assess both overt and covert behavior; we have placed it in this section on covert processes for emphasis. Self-monitoring differs from self-report in that it constitutes an observation of clinically relevant thoughts and feelings at the time of their occurrence. As such, it is a more direct method of assessment. Self-monitoring requires a child to monitor his or her own cognitions/feelings (“I’m dumb,” “I’m scared,” “I’m unhappy,” “I’m feeling out of control,” etc.) and then to record their occurrence systematically. Typically, the child is asked to keep a diary, place marks on a card, or push the plunger on a counter as cognitions/feelings occur. Although self-monitoring procedures have been used with both children and adults, at least three considerations must be attended to when such procedures are used with younger children (Shapiro, 1984). The cognitions/feelings should be clearly defined, prompts to use the procedures should be readily available, and rewards for their use should be provided. Some children will be less attuned to their inner feelings and thoughts than others and may therefore require preliminary coaching. Other children may have difficulty remembering exactly which cognitions/feelings to monitor and how those covert behaviors are defined.

For these reasons, it is generally considered desirable to provide the child a brief description of the targeted covert behaviors or, better yet, a picture of it, and to have the child record only one or two cognitions/feelings at a time. The key to successful self-monitoring in children is the use of recording procedures that are highly portable, simple, time-efficient, and relatively unobtrusive.

## Assessment of contexts

In this section we discuss methods for measuring characteristics of adult caretakers and environments which may contribute to variations in a child's overt and covert behavior. Recall that, from a social learning perspective, situational factors may exert considerable influence on a child's behavior. Methods for assessing contexts have become more plentiful as an appreciation for this influence has evolved. As might be expected, home and school environments, and the adults who interact with children in these settings, have become popular targets of assessment. In an earlier section we noted the potential value of naturalistic observation as an avenue for obtaining information about the impact of various situational factors on a child's behavior. We turn our attention now to other methods for assessing contexts.

### Interviews

Interviews with children and the adults who interact with them have the potential to yield significant information about contexts. Parents and teachers may not be highly sensitive to variations in a child's behavior in different situations and may need significant prompting along these lines ("Is Johnny aggressive *all* the time or primarily during certain activities?" "Is Susie inattentive during all class activities or especially during certain types of lessons or when certain task demands are present?" "Does Tony say he wishes he were dead only when he is in the midst of a tantrum or also at other times?"). Interviews also allow the clinician the opportunity to formally or informally assess the overt and covert behaviors of important adults in the child's life and form initial impressions about the manner in which these behaviors affect the behavior of the identified child. Children rarely refer themselves for treatment; invariably, they are referred by adults whose perceptions of a child's problems will be important to gauge. To what degree do the adults have accurate perceptions of the developmental deviance of a child's behavior? What attributions do the adults make about the child's behaviors (e.g., do they believe the behaviors are intentional, goal-oriented, or due to poor parenting, insensitive teachers, significant life events, biological factors, etc.)? Do the adults believe the child has the ability to alter behaviors defined as problematic? How do caregivers explain the situational variability of a child's challenging behaviors?

An instrument called the Assessment of Lagging Skills and Unsolved Problems (ALSUP; Greene, 2008) may be useful for assessing these variables, along with the expectations of caregivers. The ALSUP contains a list of 23 skills from the

executive, language processing, emotion regulation, cognitive flexibility, and social realms (e.g., *Difficulty expressing needs, thoughts, concerns in words; Difficulty deviating from rules, routine; Difficulty managing emotional response to frustration so as to think rationally*). Caregivers—whether parents, educators, or staff in a therapeutic facility—are engaged in a discussion aimed at identifying a child's lagging skills and the various expectations the child is having difficulty meeting (these are referred to as “unsolved problems”). Identification of unsolved problems facilitates the solving of these problems; in Greene's empirically-supported Collaborative & Proactive Solutions (CPS) model, the problem-solving is of the collaborative and proactive variety (Greene, 2014). This approach to assessment is congruent with the ideas set forth by Szasz (1961) in viewing maladaptive behavior as “problems in living.” Thus, quite intentionally, there are no normative data associated with the ALSUP, though consideration is given to whether a given skill or expectation is realistic for a given child. The ALSUP is also very much in keeping with the words of Walter Mischel some time ago (1968): “Behavioral assessment involves an exploration of the unique or idiosyncratic aspects of the single case, perhaps to a greater extent than any other approach.”

A similar instrument tapping the situational pervasiveness and severity of child behavior problems in different home settings (e.g., while playing with other children, when asked to do chores, when asked to do school homework) is the Home Situations Questionnaire (HSQ; Barkley & Edelbrock, 1987) and its revision (HSQ-R; DuPaul & Barkley, 1992). Although scores may be derived from the HSQ and HSQ-R based on the number of situations in which a child exhibits problems and the severity of these problems, these instruments are perhaps best utilized as informal devices for efficiently obtaining information about a child's behavior in a wide range of home situations. The School Situations Questionnaire (SSQ; Barkley & Edelbrock, 1987) and its revision (SSQ-R; DuPaul & Barkley, 1992) are similar to the HSQ and HSQ-R, and assess the pervasiveness of behavior problems across various school situations (e.g., during lectures to the class, during class discussions, during individual deskwork).

When it comes to intervention, assessment of parental CSLPVs is crucial. For example, one of the most commonly recommended interventions for poorly self-regulated children is behavior management strategies (e.g., contingency contracting, time-out); however, one should not assume that all parents and teachers have equal capacities (*competencies*) for implementation of such strategies, or that all parents and teachers will “stick with the program” (*self-regulatory systems and plans*) for a sufficient period (see Greene, 1995, 1996, for more thorough discussion of the issue of “teacher-treatment compatibility”). These variations in competencies and self-regulatory systems and plans may have significant ramifications for intervention selection and treatment duration. If parents have attempted such strategies in the past, it may also be fruitful to inquire about their *expectancies* regarding the likely outcome of implementing such strategies? It may also be important to assess the parents' *subjective values* as regards parenting goals and priorities. Do the parents value independence in their children? Cleanliness? Timeliness? Honesty? Good grades? Is it important to the parents that their children attend Ivy League schools?

To what degree are these subjective values compatible with the subjective values and competencies of the parents' children?

Group interviews often permit the interviewer an opportunity to observe problematic interactions (e.g., parent-child interactions, mother-father interactions, parent-teacher interactions) which may not have been included in the adults' original perception of the problematic behavior, and to explore discrepancies in reports and perceptions of all individuals present ("Mrs. Utley, your daughter seems to feel that you are constantly criticizing her... what's your sense about that?" "Mr. Jones, you don't seem as troubled by your son's behaviors as does your wife... do you feel that's an accurate perception on my part?" "I get the sense that there are very different opinions between school and home about the degree to which Adam is actually able to control his excessive motor activity").

Finally, while we have focused primarily on the covert domain in the discussion above, we wish to emphasize that interviews also provide an extremely valuable mechanism for assessing the overt behavior of important persons in the various environments in which the identified child interacts. In other words, while it is important to assess what a child's parent or teacher thinks and feels about the child, it will also be critical to assess what parents or teachers actually *do* in their interactions with the child. For example, while a parent's subjective values may suggest that she or he believes screaming is an inappropriate parenting strategy, the parent may nonetheless report that she or he screams at the child frequently. Such discrepancy between thought and action may be an important consideration in choosing efficacious treatments.

## Checklists

Various checklists have been developed to provide information about a host of contextual factors, particularly in the home and school environments. One of the most widely used global measures of the family environment is the Family Environment Scale ([Moos & Moos, 1986](#)), a parent-completed measure which provides information about family functioning in the global domains of cohesiveness, expressiveness, and conflict. A second measure has also been used to assess family cohesion and adaptability (Family Adaptability and Cohesion Evaluation Scales-II; [Olson, Bell, & Portner, 1982](#)). Other instruments may measure narrower aspects of family functioning, such as family learning environment (e.g., the Family Learning Environment Scale; [Marjoribanks, 1979](#)).

Other checklists may be useful in assessing characteristics of parents, including parental psychopathology (e.g., SCL-90-R; [Derogatis & Savitz, 2000](#)); dysfunctional discipline practices (e.g., the Parenting Scale; [Arnold, O'Leary, Wolff, & Acker, 1993](#)); the degree to which a parent finds interactions with a particular child to be stressful (e.g., the Parenting Stress Index—Fourth Edition (PSI-4; [Abidin, 2012](#)); and the degree to which parents believe they have a sound working relationship with the child's other parent (e.g., the Parenting Alliance Inventory; [Abidin & Brunner, 1995](#)).

While the above instruments are completed by parents, some instruments have been developed to assess children's perceptions of contexts, including the Social Support Scale for Children (SPPC; [Harter, 1986](#)), which assesses the degree to which a child feels supported by others (such as parents, teachers, classmates, and close friends); the Children's Report of Parenting Behavior Inventory (CRPBI; [Schluderman & Schluderman, 1970](#)), which is designed to assess a child's perceptions of his or her parents' behavior; and the Social Support Appraisals Scale (APP; [Dubow & Ullman, 1989](#)), which measures children's subjective appraisals of social support provided by family, peers, and teachers.

Other checklists may be useful in assessing characteristics of teachers. For example, the SBS Inventory of Social Behavior Standards and Expectations ([Walker & Rankin, 1983](#)) assesses teachers' behavioral expectations regarding students' behavior. Section I consists of 56 items describing adaptive student behaviors; respondents mark each item as "critical," "desirable," or "unimportant." Section II of the SBS gauges teachers' tolerances for various problematic behaviors. Section II consists of 56 items describing maladaptive student behaviors; respondents mark each item as "unacceptable," "tolerated," or "acceptable." The SBS has been shown to have excellent psychometric properties and has been used in previous studies to identify teacher expectations that are associated with effective teaching of behaviorally challenging students (e.g., [Kauffman, Lloyd, & McGee, 1989](#); [Kauffman, Wong, Lloyd, Hung, & Pullen, 1991](#)).

A second measure, the Index of Teaching Stress (ITS; [Greene & Abidin, 1995](#); [Greene, Abidin, & Kmetz, 1997](#)) assesses the degree to which a teacher experiences stress and frustration in teaching a given student, and was developed as a companion to the Parenting Stress Index described earlier. The ITS consists of two sections; the first is comprised of student behaviors which may induce stress or frustration in a teacher (this section includes five factors: ADHD, emotional ability/low adaptability, anxiety/withdrawal, low ability/learning disabled, and aggressive/conduct disorder). The second domain is comprised of various domains of teacher stress and frustration (this section includes four factors: self-doubt/needs support, loss of satisfaction from teaching, disrupted teaching process, and frustration working with parents). The utility of the ITS has been demonstrated in a longitudinal study exploring school outcome for children with ADHD ([Greene, Besztercze, Katzenstein, & Park, 2002](#)).

### ***Cultural considerations***

Numerous observers have called attention to the internationalization of the world and the "browning of America" (e.g., [Malgady, Rogler, & Constantino, 1987](#); [Vazquez Nuttall, DeLeon, & Del Valle, 1990](#); [Vazquez Nuttall, Sanchez, Borras Osorio, Nuttall, & Varvogil, 1996](#)). As regards the topic of this chapter, this means that the assessment process is increasingly being applied to non-Caucasian children for whom English is not the primary language, and that use of assessment procedures that are gender-, culture-, and language-fair has become a major concern and of utmost importance. Various cultural issues must be considered in the assessment

process. Cultural differences may be expressed in child-rearing practices, family values, parental expectations, communication styles, nonverbal communication patterns, and family structure and dynamics (Vazquez Nuttall et al., 1996).

Behaviors characteristic of ethnic minority children may be seen as emotionally or behaviorally maladaptive by persons who have little or no appreciation for cultural norms (e.g., Prewitt-Diaz, 1989). Thus, cultural biases may occur early in the referral process. Fortunately, steps can be taken to minimize cultural biases in the assessment process itself. Vazquez and colleagues (1996) have delineated a variety of steps which may be useful for incorporating cultural considerations into the assessment process: (1) including extended family members in the information-gathering process; (2) use of interpreters in interviewing; (3) familiarizing oneself with the culture of specific ethnic groups; and (4) using instruments that have been translated into the native language of reporters *and* for which norms are available for specific ethnic groups. With regard to this latter recommendation, significantly greater progress has occurred for the translation component than for the normative component. In sum, while considerable progress has been made to incorporate developmental considerations into assessment technology, the future challenge—to make similar progress in the cultural domain—is clearly before us.

## Summary

In the preceding pages we have described the various components of a behavioral assessment, reviewed these components in a historical context, and provided an overview of various assessment procedures which may be useful in conducting a thorough and comprehensive behavioral assessment. We have also asserted that perhaps the most pressing challenge for behavioral assessment at this time is the development of culturally sensitive instruments and assessors.

Yet, in outlining components of a behavioral assessment, we should emphasize several important points. First, regardless of the procedures employed, child behavioral assessments must be conducted by persons with the training and experience to execute them in a knowledgeable fashion and the skills to analyze, organize, integrate, and communicate the vast array of information gathered in the assessment process for purposes of (1) arriving at a comprehensive understanding of a child's interactions with his or her environment(s); (2) requiring that additional information be collected when such an understanding has not been achieved; (3) making accurate judgments regarding the developmental deviance of a child's behavior; (4) determining the most appropriate persons and behaviors to be targeted for change and the interventions most likely to produce these desired changes; and (5) maintaining contact over the long term with various adults who continue to interact with the child and who are charged with implementation of interventions and/or are targets of intervention; and (6) monitoring the continuous, fluid assessment process and facilitating reformulation of "the problem" as necessary. Assessors must also be well-acquainted with the nature of information provided by each assessment

procedure—in other words, what conclusions can and cannot be arrived at on the basis of the information provided by a particular instrument (Greene, 1995).

Second, while it may be obvious that, in making normative comparisons and using standardized instruments, assessors are employing certain aspects of a nomothetic approach to assessment (the application of general laws as applied to large numbers of children), we continue to view behavioral assessment of children as a primarily *idiographic* undertaking (concerned more with the uniqueness of a given child).

## References

- Abidin, R. R. (2012). *Parenting stress index 4th edition (PSI-4) manual*. Lutz, FL: Psychological Assessment Resources.
- Abidin, R. R., & Brunner, J. F. (1995). Development of a parenting alliance inventory. *Journal of Clinical Child Psychology*, 24(1), 31–40.
- Abikoff, H., & Gittelman, R. (1985). Classroom observation code: A modification of the Stony Brook code. *Psychopharmacology Bulletin*, 21(4), 901–909.
- Abikoff, H., Gittelman-Klein, R., & Klein, D. (1977). Validation of a classroom observation code for hyperactive children. *Journal of Consulting and Clinical Psychology*, 45, 772–783.
- Achenbach, T. M. (1986). *Manual for the child behavior checklist direct observation form*. Burlington: University of Vermont, Department of Psychiatry.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101, 213–232.
- Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth & Families.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.
- Arnold, D. S., O'Leary, S. G., Wolff, L. S., & Acker, M. M. (1993). The parenting scale: A measure of dysfunctional parenting in discipline situations. *Psychological Assessment*, 5, 131–136.
- Bandura, A. (1971). *Social learning theory*. Englewood Cliffs, NJ: General Learning Press.
- Bandura, A. (1973). *Aggression: A social learning analysis*. Englewood Cliffs, NJ: Prentice-Hall.
- Bandura, A. (1986). *Social foundations of thought and action: A social-cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Barkley, R. A. (1981). *Hyperactive children: A handbook for diagnosis and treatment*. New York: Guilford Press.
- Barkley, R. A. (1990). *Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment*. New York: Guilford Press.

- Barkley, R. A., & Edelbrock, C. S. (1987). Assessing situational variation in children's behavior problems: The home and school situations questionnaires. In R. Prinz (Ed.), *Advances in behavioral assessment of children and families* (Vol. 3, pp. 157–176). Greenwich, CT: JAI Press.
- Campbell, S. B. (1989). Developmental perspectives in child psychopathology. In T. H. Ollendick, & M. Hersen (Eds.), *Handbook of child psychopathology* (2nd ed). New York: Plenum Press.
- Cicchetti, D. (1984). The emergence of developmental psychopathology. *Child Development*, 55, 1–57.
- Cicchetti, D. (1993). Developmental psychopathology: Reactions, reflections, and projections. *Developmental Review*, 13, 471–502.
- Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and taxonomy. *Behavior Therapy*, 9, 882–888.
- Derogatis, L. R., & Savitz, K. L. (2000). The SCL-90-R and the brief symptom inventory (BSI) in primary care. In M. E. Maruish (Ed.), *Handbook of psychological assessment in primary care settings* (Vol. 236, pp. 297–334). Mahwah, NJ: Lawrence Erlbaum Associates.
- Diperna, J. C., & Volpe, R. J. (2005). Self-report on the social skills rating system: Analysis of reliability and validity for an elementary sample. *Psychology in the Schools*, 42(4), 345–354.
- Dubow, E. F., & Ullman, D. G. (1989). Assessing social support in elementary school children: The survey of children's social support. *Journal of Clinical Child Psychology*, 18, 52–64.
- DuPaul, G. J., & Barkley, R. A. (1992). Situational variability of attention problems: Psychometric properties of the revised home and school situations questionnaires. *Journal of Clinical Child Psychology*, 21, 178–188.
- Eyberg, S. M., Nelson, M. M., Ginn, N. C., Bhuiyan, N., & Boggs, S. R. (2013). *Dyadic parent-child interaction coding system: Comprehensive manual for research and training* (4th ed.). Gainesville, FL: PCIT International.
- Finch, A. J., Nelson, W. M., III, & Moss, J. H. (1983). A cognitive-behavioral approach to anger management with emotionally disturbed children. In A. J. Finch, W. M. Nelson, & E. S. Ott (Eds.), *Cognitive behavioral approaches to treatment with children*. Jamaica, NY: Spectrum Publications.
- Finch, A. J., & Rogers, T. R. (1984). Self-report instruments. In T. H. Ollendick, & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures*. Elmsford, NY: Pergamon Press.
- Fisher, L., Lucas, C., Sarsfield., & Shaffer. (2006). *Interviewer manual*. Center for Disease Control.
- Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *Behavior rating of executive function*. Lutz, FL: Psychological Assessment Resources.
- Greene, R. W. (1995). Students with ADHD in school classrooms: Teacher factors related to compatibility, assessment, and intervention. *School Psychology Review*, 24(1), 81–93.
- Greene, R. W. (1996). Students with ADHD and their teachers: Implications of a goodness-of-fit perspective. In T. H. Ollendick, & R. J. Prinz (Eds.), *Advances in clinical child psychology* (pp. 205–230). New York: Plenum.
- Greene, R. W. (2008). *Lost at school: Why our kids with behavioral challenges are falling through the cracks and how we can help them*. New York: Scribner.
- Greene, R. W. (2014). *The explosive child: A new approach for understanding and parenting easily frustrated, chronically inflexible children*. New York: HarperCollins.

- Greene, R.W., & Abidin, R.R. (1995). The index of teaching stress: A new measure of student teacher compatibility. In Paper presented at the 27th annual meeting of the National Association of School Psychologists, Chicago, IL.
- Greene, R. W., Abidin, R. R., & Kmetz, C. (1997). The index of teaching stress: A measure of student–teacher compatibility. *Journal of School Psychology*, 35(3), 239–259.
- Greene, R. W., Besztercze, S. K., Katzenstein, T., & Park, K. (2002). Are students with ADHD more stressful to teach? Patterns and predictors of teacher stress in an elementary-age sample. *Journal of Emotional and Behavioral Disorders*, 10, 27–37.
- Greenwood, C. R., Carta, J. J., Kamps, D., Terry, B., & Delquadri, J. (1994). Development and validation of standard classroom observation system for school practitioners: Ecobehavioral assessment software EBASS. *Exceptional Children*, 61, 197–210.
- Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system manual*. Circle Pines, MN: American Guidance Service.
- Gross, A. M. (1984). Behavioral interviewing. In T. H. Ollendick, & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures*. Elmsford, NY: Pergamon Press.
- Harter, S. (1985). *The self-perception profile for children: Revision of the perceived competence scale for children [manual]*. Denver, CO: University of Colorado.
- Harter, S. (1986). *Manual: Social support scale for children*. Denver, CO: University of Denver.
- Harter, S. (1988). *Manual for the self-perception profile for adolescents*. Denver, CO: University of Denver 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.
- Hops, H., Biglan, A., Tolman, A., Sherman, L., Arthur, J., Warner, P., ... Osteen, I. (1990). *Living in familial environments (UFE) coding system: Training/procedures and reference manual for coders (rev. ed.)*. Eugene: Oregon Research Institute.
- Jacobs, J. R., Boggs, S. R., Eyberg, S. M., Edwards, D., Durning, P., Querido, J. G., ... Funderburk, B. W. (2000). Psychometric properties and reference point data for the revised edition of the school observation coding system. *Behavior Therapy*, 31, 695–712.
- Johnson, S. M., & Bolstad, O. D. (1973). Methodological issues in naturalistic observations: Some problems and solutions for field research. In L. A. Harnmerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice*. Champaign, IL: Research Press.
- Kauffman, J. M., Lloyd, J. W., & McGee, K. A. (1989). Adaptive and maladaptive behavior: Teachers' attitudes and their technical assistance needs. *Journal of Special Education*, 23, 185–200.
- Kauffman, J. M., Wong, K. L. H., Lloyd, J. W., Hung, L. Y., & Pullen, P. L. (1991). What puts pupils at risk? An analysis of classroom teachers' judgments of pupils' behavior. *Remedial and Special Education*, 12, 7–16.
- Korkman, M., Kirk, U., & Kemp, S. L. (2007a). *NEPSY II. Administrative manual*. San Antonio, TX: Psychological Corporation.
- Korkman, M., Kirk, U., & Kemp, S. L. (2007b). *NEPSY II. Clinical and interpretative manual*. San Antonio, TX: Psychological Corporation.
- Kovacs, M. (2014). *Children's depression inventory-2*. Los Angeles: Multi-Health Systems.
- LaGreca, A. M., & Stone, W. L. (1993). Social anxiety scale for children-revised: Factor structure and concurrent validity. *Journal of Clinical Child Psychology*, 22, 17–27.

- Linehan, M. (1977). Issues in behavioral interviewing. In J. D. Cone, & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel.
- Malgady, R., Rogler, L., & Constantino, G. (1987). Ethnocultural and linguistic bias in mental health evaluation of Hispanics. *American Psychologist*, 42, 228–234.
- Marjoribanks, K. (1979). *Families and their learning environments*. London: Routledge & Kegan Paul.
- Mash, E. J., & Barkley, R. A. (1986). Assessment of family interaction with the response class matrix. In R. Prinz (Ed.), *Advances in behavioral assessment of children and families* (Vol. 2, pp. 29–67). Greenwich, CT: JAI Press.
- Mash, E. J., & Terdal, L. G. (1988). Behavioral assessment of child and family disturbance. In E. J. Mash, & L. G. Terdal (Eds.), *Behavioral assessment of childhood disorders* (pp. 3–65). New York: Guilford Press.
- McConaughy, S. H. (1996). The interview process. In M. J. Breen, & C. R. Fiedler (Eds.), *Behavioral approach to assessment of youth with emotional/behavioral disorders: A handbook for school-based practitioners* (pp. 181–224). Austin, TX: ProEd.
- McConaughy, S. H., & Achenbach, T. M. (2001). *Manual for the semistructured clinical interview for children and adolescents* (2nd ed). Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.
- McConaughy, S. H., Achenbach, T. M., & Gent, C. L. (1988). Multiaxial empirically based assessment: Parent, teacher, observational, cognitive, and personality correlates of child behavior profiles for 6- to 11-year-old boys. *Journal of Abnormal Child Psychology*, 16, 485–509.
- McLeod, B. D., Jensen-Doss, A., & Ollendick, T. H. (Eds.). (2013). *Diagnostic and behavioral assessment: A clinical guide*. New York: The Guilford Press.
- Mischel, W. (1968). *Personality and assessment*. New York: Wiley.
- Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, 80, 252–283.
- Mischel, W. (1979). On the interface of cognition and personality. *American Psychologist*, 34, 740–754.
- Mischel, W. (1984). Convergences and challenges in the search for consistency. *American Psychologist*, 39, 351–364.
- Moos, R. H., & Moos, B. S. (1986). *Family environment scale manual* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.
- Muris, P., Ollendick, T. H., Roelofs, J., & Austin, K. (2014). The short form of the fear survey schedule for children-revised (FSSC-R-SF). *Journal of Anxiety Disorders*, 28, 957–965.
- Novick, J., Rosenfeld, E., Bloch, D. A., & Dawson, D. (1966). Ascertaining deviant behavior in children. *Journal of Consulting and Clinical Psychology*, 30, 230–238.
- Nowicki, S., & Strickland, B. R. (1973). A locus of control scale for children. *Journal of Consulting and Clinical Psychology*, 40, 148–154.
- Ollendick, T. H. (1983). Reliability and validity of the revised-fear survey schedule for children (FSSC-R). *Behaviour Research and Therapy*, 21, 685–692.
- Ollendick, T. H., & Cerny, J. A. (1981). *Clinical behavior therapy with children*. New York: Plenum Press.
- Ollendick, T. H., & Greene, R. W. (1990). Behavioral assessment of children. In G. Goldstein, & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd ed). New York: Pergamon.

- Ollendick, T. H., & Hersen, M. (Eds.), (1984). *Child behavioral assessment: Principles and procedures*. New York: Pergamon Press.
- Ollendick, T. H., King, N. J., & Frary, R. B. (1989). Fears in children and adolescents: Reliability and generalizability across gender, age, and nationality. *Behaviour Research and Therapy*, 27, 19–26.
- Ollendick, T. H., Matson, J. L., & Helsel, W. J. (1985). Fears in children and adolescents: Normative data. *Behaviour Research and Therapy*, 23, 465–467.
- Ollendick, T. H., Yang, B., King, N. J., Dong, Q., & Akande, A. (1996). Fears in American, Australian, Chinese, and Nigerian children and adolescents: A cross-cultural study. *Journal of Child Psychology and Psychiatry*, 37, 213–220.
- Olson, D. H., Bell, R. O., & Portner, J. (1982). *Manual for FACES II: Family adaptability and cohesion scales*. St. Paul, MN: University of Minnesota, Family Social Science.
- Patterson, G. R. (1976). The aggressive child: Victim and architect of a coercive system. In E. J. Mash, L. A. Hammerlynck, & L. C. Hardy (Eds.), *Behavior modification and families*. New York: Brunner/Mazel.
- Piers, E. V. (1984). *Revised manual for the Piers Harris children's self-concept scale*. Los Angeles: Western Psychological Services.
- Prewitt-Diaz, J. (1989). *The process and procedures for identifying exceptional language minority children*. State College: Pennsylvania State University.
- Reynolds, C. R., & Richmond, B. O. (2008). *RCMAS-2: revised children's manifest anxiety scale: Second edition [manual]*. Los Angeles, CA: Western Psychological Services.
- Rotter, J. B. (1954). *Social learning and clinical psychology*. Englewood Cliffs, NJ: Prentice-Hall.
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 80. (Whole No. 609).
- Rotter, J. B. (1972). Beliefs, social attitudes, and behavior: A social learning analysis. In J. B. Rotter, J. E. Chance, & E. J. Phares (Eds.), *Applications of a social learning theory of personality*. New York: Holt, Rinehart, & Winston.
- Rutter, M., & Garmezy, N. (1983). Developmental psychopathology. In P. Mussen (Ed.), *Handbook of child psychopathology* (Vol. 4, pp. 775–911). New York: Wiley.
- Sameroff, A. J., & Chandler, M. J. (1975). Reproductive risk and the continuum of caretaking casualty. In F. B. Horowitz (Ed.), *Review of child development research* (4, pp. 187–244). Chicago: University of Chicago Press.
- Scherer, M. W., & Nakamura, C. Y. (1968). A fear survey schedule for children (FSS-FC): A factor-analytic comparison with manifest anxiety (CMAS). *Behaviour Research and Therapy*, 6, 173–182.
- Schluderman, E., & Schluderman, S. (1970). Replicability of factors in children's report of parent behavior (CRPBI). *Journal of Psychology*, 76, 239–249.
- Shapiro, E. S. (1984). Self-monitoring. In T. H. Ollendick, & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures*. Elmsford, NY: Pergamon Press.
- Silverman, W. K., & Albano, A. M. (1996). *The anxiety disorders interview schedule for children for DSM-IV (child and parent versions)*. San Antonio, TX: Psychological Corporation.
- Silverman, W. K., Fleisig, W., Rabian, B., & Peter son, R. A. (1991). Childhood anxiety sensitivity index. *Journal of Clinical Child Psychology*, 20, 162–168.
- Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.
- Spielberger, C. D. (1973). *State-trait anxiety inventory for children: Preliminary manual*. Palo Alto, CA: Consulting Psychologists Press.

- Swann, G. E., & MacDonald, M. L. (1978). Behavior therapy in practice: A rational survey of behavior therapists. *Behavior Therapy*, 9, 799–807.
- Szasz, Thomas S. (1961). *The myth of mental illness: Foundations of a theory of personal conduct*. New York: Harper & Row.
- Vazquez Nuttall, E., DeLeon, B., & Del Valle, M. (1990). Best practice in considering cultural factors. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology II* (pp. 219–233). Washington, DC: National Association of School Psychologists.
- Vazquez Nuttall, E., Sanchez, W., Borras Osorio, L., Nuttall, R. L., & Varvogil, L. (1996). Assessing the culturally and linguistically different child with emotional and behavioral problems. In M. J. Breen, & C. R. Fiedler (Eds.), *Behavioral approach to assessment of youth with emotional/behavioral disorders: A handbook for school based practitioners* (pp. 451–502). Austin, TX: ProEd.
- Wahler, R. G. (1976). Deviant child behavior in the family: Developmental speculations and behavior change strategies. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, NJ: Prentice-Hall.
- Walker, H. M., & Rankin, R. (1983). Assessing the behavioral expectations and demands of less restrictive settings. *School Psychology Review*, 12, 274–284.
- Wechsler, D. (2014). *Wechsler intelligence scale for children—Fifth edition*. Bloomington, MN: Pearson.
- Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical evaluation of language fundamentals—Fifth edition (CELF-5)*. Bloomington, MN: NCS Pearson.
- Winett, R. A., Riley, A. W., King, A. C., & Altman, D. G. (1989). Preventive strategies with children and families. In T. H. Ollendick, & M. Hersen (Eds.), *Handbook of child psychopathology* (2nd ed). New York: Plenum Press.
- World Health Organization. (2013). *ICD-10-CM international classification of diseases, tenth revision, clinical modification*. Geneva: Author.

## Further reading

- Gersten, R., Walker, H., & Darch, C. (1988). Relationship between teachers' effectiveness and their tolerance for handicapped students. *Exceptional Children*, 54, 433–438.
- Scanlon, E. M., & Ollendick, T. H. (1985). Children's assertive behavior: The reliability and validity of three self-report measures. *Child and Family Behavior Therapy*, 7, 9–21.

# Behavioral assessment of adults in clinical settings

16

Stephen N. Haynes<sup>1</sup>, William H. O'Brien<sup>2</sup> and Joseph Keawe'aimoku Kaholokula<sup>3</sup>

<sup>1</sup>Department of Psychology, University of Hawai'i at Mānoa, Honolulu, HI, United States,

<sup>2</sup>Department of Psychology, Bowling Green State University, Bowling Green, OH, United States, <sup>3</sup>Department of Native Hawaiian Health, John A. Burns School of Medicine, University of Hawai'i at Mānoa, Honolulu, HI, United States

## Behavioral assessment with adults in clinical settings

### ***Measurement and clinical science in psychological assessment***

There are three primary purposes of psychological assessment in clinical settings. The first purpose is to identify, operationally define, and measure a client's adaptive and maladaptive behaviors and treatment goals. A second purpose is to identify, operationally define, and measure factors that influence a client's adaptive and maladaptive behaviors and attainment of treatment goals. A third purpose is to integrate assessment information in order to design and evaluate interventions that can improve a client's quality of life. Consider, for example, the challenging assessment tasks confronting a psychologist with a client who is experiencing intense panic, social isolation, and frequent conflicts with a partner. First, the clinician must select an assessment strategy that will effectively capture and evaluate these multiple problems across the course of the intervention. Further, the assessment strategy must enable the clinician to identify important causal relations associated with these problems in order to understand why the client is experiencing isolation, panic episodes, and conflict. Finally, the assessment strategy and resultant information must be synthesized and then used to design an intervention that will modify causal relations in order to promote a reduction in panic episodes, isolation, and conflict and simultaneously promote an increase in adaptive behaviors. The aforementioned assessment goals exemplify the clinical and research applications of psychological assessment—the systematic measurement of a person's behavior, variables associated with variance in behavior, and the inferences and judgments based on those measures (see multiple definitions of psychological assessment in Geisinger, 2013; Haynes, Smith, & Hunsley, 2019). We use the term “behavior” to include overt actions, emotions, cognitive processes, and physiological responses. Additionally, the term “variables” includes behavioral, environmental, social, and biological variables.

Psychological assessment paradigms differ in their fundamental assumptions, applicability, utility, and preferred assessment strategies. A psychological assessment paradigm refers to the assumptions, beliefs, values, hypotheses, and methods endorsed within an assessment discipline.<sup>1</sup> All psychological assessment paradigms are at least partially explanatory. That is, they are designed to elucidate the causes of human behavior. For example, some psychodynamic paradigms may assume that the panic, social isolation, and conflict of the aforementioned client significantly arises from historical, developmental, unconscious, and intrapsychic processes. Within this paradigm it is assumed that these causes can be best identified using the client's verbal reports of perceptions when asked to view ambiguous stimuli, such as a Rorschach or Thematic Apperception Test. Some personality-based paradigms assume that the client's problems often result from temporally and situationally consistent patterns of cognitive, emotional, and behavioral dispositions that can be identified through the person's responses to items on a self-report symptom inventory. Edited books by Hersen (2006) and Geisinger (2013), and other chapters in this book, provide overviews of the conceptual foundations and assessment strategies associated with various psychological assessment paradigms and subparadigms.

This chapter focuses on clinical applications of behavioral assessment with adults for the purposes of identifying behavior problems, treatment goals, and factors that affect them. The ultimate purpose of behavioral assessment in a clinical context is to facilitate the design of a maximally effective intervention for a client. Psychological assessment, and the behavioral assessment paradigm, also play an important role in forensic evaluations (Hart, Gresswell, & Braham, 2011; Heilbrun, DeMatteo, Brooks Holliday, & LaDuke, 2014), parenting competence evaluations in custody, abuse, and neglect cases (Budd, Connell, & Clark, 2011), neuropsychological assessment with traumatic brain injury (Granacher, 2015), personnel selection and evaluation (Guion, 2011; Hough & Oswald, 2000), psychiatric diagnosis (American Psychiatric Association, 2013), and couples and family therapy (Snyder, Heyman, Haynes, Carlson, & Balderrama-Durbin, 2016).

All psychological assessment paradigms involve the measurement of behavior. Although information from psychological assessment is partially qualitative, such as in a clinician's judgments about a client drawn from interviews, most psychological assessment paradigms also involve the assignment of quantitative values (i.e., measures/scores) to dimensions of a person's behavior. Dimensions are quantitative aspects of a phenomenon and can include the severity and duration of depressed moods, the amount of alcohol consumed in a week, the level and variability of resting diastolic blood pressure, the frequency and recency of lifetime traumatic life events experienced by a person, the likelihood of a positive contingency for behavior in a classroom, or the onset latency of sleep at bedtime.

<sup>1</sup> Major psychological assessment paradigms often include subparadigms that differ in their emphases on types of variables and methods. For example, a behavioral assessment paradigm includes behavior analytic, cognitive-behavioral, and social learning paradigms.

In contrast to other psychological assessment paradigms, behavioral assessment emphasizes the measurement of behavior in context (Eckert & Lovett, 2013; Haynes, O'Brien, & Kaholokula, 2011; O'Brien, Kaholokula, & Haynes, 2016). Additionally, the strength of functional relations between behavior and contextual variables is an important quantitative dimension for designing interventions in behavioral assessment. This emphasis on quantitative measurement is a foundation of behavioral assessment in the clinical sciences.<sup>2</sup> Without precise measurement, hypothesized functional relations for behavior problems cannot be identified and tested; intervention effects cannot be evaluated; the precision, utility, and applicability of an assessment and intervention paradigm cannot evolve; and clients will not receive the best services.

Behavioral assessment also emphasizes quantification because it strengthens the precision, validity, and utility of clinical judgments and inferences derived from nomothetic research. Measures that have insufficient psychometric quality (e.g., those with poor construct validity, precision, or sensitivity) can lead to erroneous and sometimes harmful judgments about persons seeking treatment for behavior problems. Poor quality measures can also increase the likelihood that a clinician will make errors in clinical judgment, such as misidentifying a client's problems and treatment goals, drawing incorrect inferences about the causes of behavior, designing ineffective interventions, and inaccurately estimating the effectiveness of interventions.

As discussed in other sources (Fisher, O'Donohue, & Haynes, 2018; Haynes et al., 2019), the psychometric properties of measures from an assessment instrument set upper limits on their utility for clinical judgments. Estimates of causal relations in clinical assessment are particularly important because many interventions attempt to modify causal variables that are judged to exert important influences on behavior problems or goal attainment. Empirically based and quantitatively focused measurement strategies enable the assessor to more adequately operationalize a client's behavior problems and goals, predict the client's future behavior, draw inferences about the factors that affect the client's behavior problems, and identify potential moderators of treatment goal attainment. In turn, these judgments help the clinician to select the best intervention for the client and to evaluate its process, time course, and effectiveness.

## ***Overview of the chapter***

The goals of this chapter are to describe and advocate for psychological assessment strategies that: (1) use the best available evidence-based measurement and clinical judgment strategies; (2) reflect the context dependent, conditional, and multimodal

<sup>2</sup> Controversies surrounding the theory, foundations, models, meaning, and units of measurement in psychology are important but outside the domain of this chapter. Readers can consult presentations on this topic by McGrane (2015) and Howell, Breivik, and Wilcox (2007). In this chapter, we emphasize the use of the *best* available science-based measurement strategies and acknowledge that all measures partially reflect error.

nature of behavior; (3) attend to the functional relations associated with behavior change; (4) are sensitive to diversity and individual differences among clients; (5) capture to the dynamic nature of behavior and functional relations; and (6) promote direct and minimally inferential approaches to measurement. We emphasize the conceptual foundations of behavioral assessment and their influence on behavioral assessment strategies and methods, particularly the dynamic, conditional, multimodal, and idiographic aspects of behavioral problems, goals, and causal relations. We also highlight the utility of integrating a functional analysis approach to assessment and clinical case formulations. We conclude the chapter with a set of recommendations for conducting psychological assessment in clinical contexts.

Because of space limitations, this chapter focuses on clinical assessment with adults. However, the conceptual foundations, principles, strategies, and methods discussed herein are also relevant to clinical assessment, research, and program evaluation with children. Extended discussions of behavioral assessment broadly applied can be found in [Haynes, O'Brien, et al. \(2011\)](#) and [O'Brien et al. \(2016\)](#). Psychometric principles of clinical assessment have been summarized in [Haynes et al. \(2019\)](#) and behavioral assessment with an emphasis on children is discussed in [Eckert and Lovett \(2013\)](#), [McLeod, Jensen-Doss, and Ollendick \(2013\)](#) and other chapters in this book. [Kazdin \(2001\)](#) and [O'Brien et al. \(2016\)](#) review some of the learning theory foundations of behavioral assessment. Behavioral case formulations and functional analysis are discussed in [Haynes et al. \(2019\)](#), [O'Brien et al. \(2016\)](#), [Persons \(2008\)](#), and [Sturmey \(2008\)](#).

### *A note on idiographic and nomothetic assessment strategies*

This chapter focuses on behavioral assessment as both an *idiographic* and *nomothetic* assessment strategy. In an idiographic assessment strategy, the measures obtained are interpreted through reference to other measures from the same person obtained at different times or in different contexts ([Haynes, Mumma, & Pinson, 2009](#)). An idiographic assessment of the client described earlier might focus on monitoring changes in the rate, severity, or duration of the client's panic episodes across time and contexts using ecological momentary assessment (EMA, [Ebner-Priemer & Trull, 2009](#)). Clinical judgments about the client's panic episodes could be based on daily reports of panic and the presence or absence of key factors, such as social stressors or generalized anxiety, associated with those changes. In these examples, the frame of reference for interpreting assessment data is within-person.

In a nomothetic approach to assessment, the measures obtained in clinical assessment are interpreted in reference to data obtained from other persons. Essentially, the frame of reference for interpreting assessment data is between-persons. For example, we could compare the client's scores on a standardized assessment instrument, such as the Beck Anxiety Inventory ([Beck & Steer, 1990](#)), to scores from others in order to evaluate the relative severity of the client's anxiety and panic episodes. As we discuss in the section on strategies of behavioral assessment, idiographic and nomothetic strategies can be integrated.

## ***Conceptual foundations of behavioral assessment***

Clinical assessment strategies are guided by assessment paradigms, which are assumptions about the nature of human behavior and the variables and processes that influence it. Behavioral assessment is conceptually driven and methodologically ecumenical. That is, it includes a diverse set of assessment methods and strategies that are consistent with certain evolving premises about behavior. Additionally, behavioral assessment promotes a functional and idiographic approach to assessment based on the presumption that the best clinical assessment strategy will differ across clients because of their unique characteristics, culture, life contexts, and goals of the assessment.

Behavioral assessment is above all a science-based approach to psychological assessment. Although our understanding of human behavior is evolving, we know that clinical assessment must be congruent with findings from studies that have documented the multimodal (e.g., overt behavior, thoughts, physiological reactions), dynamic (changes over time), conditional and contextual (variations across settings and events), and idiographic nature of behavior. Based on this evidence, behavioral assessment emphasizes the importance of basing clinical judgments on precise (i.e., valid and sensitive to change) measures obtained across multiple response modes, contexts, and sources of information. Behavioral assessment also encourages the acquisition of direct and minimally inferential measures that are appropriate for the cultural context of the individual and the goals of assessment.<sup>3</sup>

Tables 16.1–16.3 outline the tenets that guide behavioral assessment strategies, and illustrate how they encourage the use of assessment strategies that are appropriately validated and include multiple sources, methods, situations, contexts, times, and dimensions. In these tables, we highlight how behavioral assessment strategies are guided by research findings indicating that clients often have multiple behavior problems that (1) co-occur, (2) are functionally interrelated, (3) vary in characteristics across clients, settings, conditions, contexts, and time, and (4) have multiple attributes, response modes, and dimensions that differ in importance. We also note the multiple, idiographic, bidirectional, interactive, and moderated and mediated nature of causal variables and relations [see reviews of psychopathology in [Maddux and Winstead \(2012\)](#); and scientific foundations of clinical assessment principles in [Haynes et al. \(2018\)](#)]. A more detailed discussion of these strategies is provided in a later section of this chapter under “Strategies of Behavioral Assessment.”

A particularly important implication of the idiographic, conditional, interactive, and dynamic nature of behavior problems is that broadly focused (i.e., aggregated), single snap-shot-time-sampled, descriptive and diagnostically oriented, nonconditional assessment strategies, such as those associated with personality, projective, neuropsychological, and cognitive assessments ([Geisinger, 2013](#)), can be useful but provide an insufficient basis for generating optimally effective intervention

<sup>3</sup> Consistent with the definition in [Haynes, Kaholokula, and Tanaka-Matsumi \(2018\)](#), we define “culture” as “the shared patterns of behaviors and interactions, cognitive constructs, and affects that are learned through a process of socialization and that distinguish members of a group from members of another group.”

**Table 16.1** Conceptual foundations of behavioral assessment: the nature of behavior problems

<b>The nature of behavior problems</b>	<b>Description/elaboration/example</b>	<b>Implications for behavioral assessment strategies and methods</b>
<p><i>Multiple behavior problems:</i> Clients can have multiple co-occurring and interacting behavior problems that differ in importance.</p>	<p>Clients who present with depressed mood often have co-occurring and interacting anxiety, interpersonal, and/or substance use problems.</p>	<p>Establish the relative importance and functional relations among the problem behaviors; delay intervention decisions until the functional relations among multiple problems have been evaluated.</p>
<p><i>Multiple response modes:</i> Behavior problems can be expressed in cognitive, behavioral, physiological, and emotional response modes that may or may not strongly covary and can differ in importance and causal relations across clients.</p>	<p>Clients who present with anxiety symptoms can differ in the relative importance of behavioral, cognitive, emotional, and physiological symptoms and the variables that influence them.</p>	<p>Identify the most important response modes for a client so that causal relations can be identified and intervention effects can be more sensitively monitored.</p>
<p><i>Multiple dimensions:</i> Behavior problems can have multiple dimensions (e.g., severity, duration) that differ in their importance and associated causal relations.</p>	<p>Clients can differ in the importance of frequency, duration, and severity of manic episodes and each dimension can be influenced by different causal relations.</p>	<p>Identify the most important dimension of a client's behavior problem so that causal relations can be identified and treatment can be optimally directed and monitored.</p>
<p><i>Contextual nature:</i> The aspects and dimensions of behavior problems can be conditional and context dependent.</p>	<p>The likelihood that a client's symptoms of PTSD (e.g., re-experiencing) will occur can depend on concurrent life stressors and social support.</p>	<p>Identify the contexts associated with the differential likelihood of occurrence of a behavior problem (or another associated dimension) so that causal relations can be identified and treatment effects are sensitively monitored.</p>

(Continued)

**Table 16.1** (Continued)

<b>The nature of behavior problems</b>	<b>Description/elaboration/example</b>	<b>Implications for behavioral assessment strategies and methods</b>
<i>Dynamic nature:</i> The characteristics, dimensions, and causal relations associated with behavior problems can be dynamic and unstable.	The form, rate, and causal relations associated with a client's alcohol and drug use can change over time.	Behavior problems and the variables that influence them should be frequently monitored during intervention and follow-up to evaluate and quickly identify clinically relevant changes.
<i>Heterogeneous nature:</i> Formal diagnostic categories (e.g., DSM-5) are composed of heterogeneous sets of symptoms that often do not co-occur to a high degree and can be affected by different causal variables.	Clients with a diagnosis of Major Depressive Disorder can differ in the degree to which they experience the nine major symptoms listed in DSM-5 and can differ in the causal relations associated with each symptom.	Diagnosis and problem identification can be helpful but are insufficient to adequately inform intervention decisions; additional data are needed on the component symptoms and causal relations associated with them.

**Table 16.2** Conceptual foundation of behavioral assessment: the nature of causal relations with behavior problems

<b>The nature of causal relations associated with behavior problems</b>	<b>Description/elaboration/example</b>	<b>Implications for behavioral assessment strategies and methods</b>
<i>Multiple causality:</i> A behavior problem can be influenced by different permutations of <i>multiple causal variables</i> and these causal permutations can vary across persons with the same behavior problem.	Clients can differ in the degree to which their manic episodes are affected by sleep impairment, medication, positive experiences, and rebound from a depressive state.	Examine the causal role of all potential important causal variables; avoid premature assumptions about the causal influences for a client's behavior problem; avoid premature intervention decisions.
<i>Multiple attributes:</i> Causal variables for behavior problems can have <i>multiple attributes</i> that	Clients differ in the degree to which symptoms of trauma-related distress are associated with the	Identify the specific aspects of causal variables that are most

(Continued)

**Table 16.2** (Continued)

The nature of causal relations associated with behavior problems	Description/elaboration/example	Implications for behavioral assessment strategies and methods
differ in their degree of influence on a client's behavior problem.	severity, duration, sights, sounds, or smells associated with the traumatic event.	strongly affect a behavior problem.
<p><i>Interactive causal paths:</i> Causal variables for behavior problems can <i>interact</i>, form <i>chains</i>, and influence behavior problems through <i>multiple causal paths</i>.</p>	A client's stressful work experience can affect sleep quality and both can affect the client's experience of pain, independently and in interaction.	Identify the potential directions of causal relations among causal variables and behavior problems and how they affect a behavior problem independently and in combination.
<p><i>Causal directionality:</i> Causal variables and behavior problems can evidence <i>bidirectional causal relations</i>.</p>	Conflict with an intimate partner can lead to an increase in the use of alcohol/drugs, which, in turn, can increase the chance of intimate partner conflict.	Identify the form (e.g., unidirectional, bidirectional) of functional relations among causal variables and behavior problems.
<p><i>Dynamic causal relations:</i> Causal variables associated with a behavior problem can be <i>dynamic and unstable</i>—they can change over time.</p>	Alcohol or drug use can initially be more strongly associated with social factors and after continued use more strongly associated with biological factors.	Periodically examine causal relations for a client's behavior problem throughout the assessment/intervention process.
<p><i>Moderator variables</i> can strongly affect causal relations associated with a behavior problem.</p>	The likelihood that a client will experience depressive symptoms following a stressor can be affected by the client's level of social support from friends and family.	Identify variables that could moderate the effects of an important causal relation for a client's behavior problem or treatment goal attainment.
<p><i>Mediators and causal mediation</i> can account for or explain causal relations associated with a client's behavior problem.</p>	The effect of a distressing social experience on a client's anxiety symptoms can be mediated by the client's interpretation of physiological responses during the experience.	During the assessment process consider "why," "how," or "what accounts for" an identified causal relation associated with a client's behavior problem.

(Continued)

**Table 16.2** (Continued)

The nature of causal relations associated with behavior problems	Description/elaboration/example	Implications for behavioral assessment strategies and methods
<p><i>Contemporaneous causal relations</i> can have particularly important implications for intervention.</p>	<p>Early childhood abuse can be a significant causal factor for a client's substance use but contemporaneous thoughts, emotions, and behavioral skills associated with the abuse can be very useful treatment foci.</p>	<p>While identifying the causal role of historical events and prior learning, attend especially to the contemporaneous behavioral, cognitive, emotional, and physiological functional relations associated with behavior problems.</p>
<p>Causal relations can be <i>conditional</i> and <i>context dependent</i>.</p>	<p>The factors that influence a client's bulimic or binge eating can vary with the social setting, biological state, or recent experiences.</p>	<p>Attend to the importance of the conditional nature of causal relations and evaluate causal relations within different contexts.</p>
<p><i>Behavior-environment interactions</i> can be particularly important causal influences for a client's behavior problems (i.e., <i>reciprocal determinism</i>).</p>	<p>The ways in which a client interacts with others can strongly affect his or her social anxiety, self-image, mood, and social support.</p>	<p>Attend to the ways in which the client interacts with others in order to identify behavior changes that might benefit the client; examine a client's behavioral skills and skills deficits.</p>
<p><i>Response contingencies</i> can be a particularly important aspect of the behavior-environment interactions that influence a client's behavior problem.</p>	<p>The responses of staff members can affect the likelihood that a hospitalized patient will exhibit appropriate or inappropriate social behaviors.</p>	<p>Attend to the ways in which persons in the client's environment respond to specific positive and negative behaviors.</p>
<p><i>Extended social systems</i> can influence a client's behavior problems.</p>	<p>A client's excessive dieting can be influenced by interactions with friends and family.</p>	<p>Consider the client's extended social environment (friends, family, coworkers, teachers) when evaluating potential causal variables for the client's behavior problems.</p>

**Table 16.3** Conceptual foundation of behavioral assessment: psychometric principles of measurement and clinical assessment

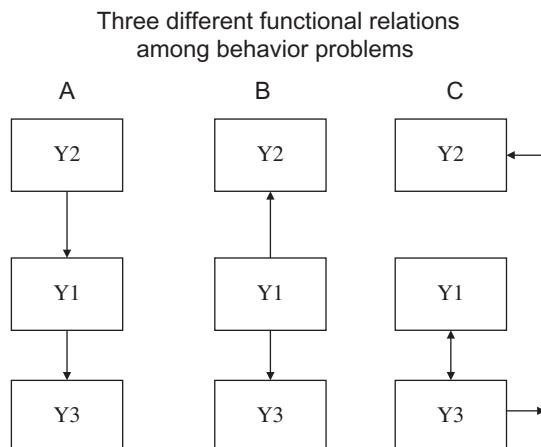
Psychometric principles	Description/elaboration/example	Implications for behavioral assessment strategies and methods
<p>Clinical judgments about a client should be based on measures with <i>sound and assessment-context relevant psychometric evidence</i>.</p>	<p>Judgments about a client's depressive symptoms and the factors that influence it should be based on measures with strong psychometric evidence.</p>	<p>Select measures that have strong psychometric support for the assessment contexts in which they will be used.</p>
<p><i>Validity and accuracy</i> are often more important than reliability.</p>	<p>A measure of social anxiety can show high levels of internal consistency and test-retest reliability but still be a poor measure of social anxiety.</p>	<p>Select measures that have been subjected to strong evaluations of convergent, discriminant, and content validity.</p>
<p>Measures used in clinical assessment should be <i>appropriate for the culture of the client</i>.</p>	<p>Instruments should provide measures that have been shown to be valid given the client's ethnicity, age, sexual orientation, and other dimensions of individual difference.</p>	<p>Select measures with strong psychometric support from individuals similar to the client on dimensions of individual differences and diversity.</p>
<p>An assessment instrument can provide <i>multiple measures</i> that can differ in their psychometric support.</p>	<p>An instrument to measure social anxiety can provide scale scores on cognitive, behavioral, and emotional response modes that can differ in their level of psychometric support.</p>	<p>Do not ascribe psychometric evidence to an assessment instrument with multiple measures; use only measures from an instrument that have been strong psychometric support.</p>
<p>Psychometric evidence for a measure can vary as a function of <i>the assessment context and goal</i>.</p>	<p>A measure to identify early signs of psychosis can be valid and useful as a brief screen but less so for clinical case formulation.</p>	<p>Consider the assessment purpose and context when evaluating the psychometric evidence for a measure.</p>
<p><i>Sources of measurement error</i> differ across methods of assessment, assessment instruments, and sources.</p>	<p>Self-report and analog observation measures of intimate partner conflict can differ in their bias and reactive effects.</p>	<p>Multimethod and multisource assessment can provide more comprehensive measures of behavior and functional relations.</p>

decisions. Similarly, a diagnosis does not capture important dynamic, conditional, and specific aspects of a client's behavior problems. A diagnosis is also insufficiently sensitive to individual differences across clients in clinically important functional relations associated with their behavior problems and treatment goals, cannot provide a sensitive measure of change across treatment sessions, and is an insufficient basis for intervention design.

For our client presented earlier, a score on an anxiety inventory or a diagnosis of panic disorder would not help us understand the unique and specific aspects of the client's responses to environmental events or the factors that influence day-to-day fluctuation in anxiety and panic episodes. Data relevant to these unique and specific behavior problem characteristics can best be derived from more focused assessment strategies, such as functional behavioral interviews, specific and functionally focused questionnaires, cognitive assessments, and rating scales, psychophysiological measures, systematic analog and natural environment observation, and EMA in the client's daily life. Data from personality tests can indicate the likelihood of particular aspects of behavior problems, but provide little information about their clinically important functional relations.

Consider the well-documented high rates of comorbidity in clinical samples (e.g., Krueger & Markon, 2006) and the importance of identifying functional relations among a client's behavior problems. The optimal treatment focus would differ according to patterns of functional relations between problems. For example, multiple forms of causal relations could be associated with a client's panic, social isolation, and relationship conflict. In turn, the different causal relations would result in different intervention foci.

This point is illustrated in Fig. 16.1 where Y1 indicates panic and anxiety, Y2 indicates intimate partner conflict, and Y3 indicates social isolation. In Functional



**Figure 16.1** Different causal relations (3 of potentially 12) among three behavior problems to illustrate that different patterns of causal relations would lead to different decisions about the best treatment foci for different clients. Y1 indicates panic or anxiety episodes, Y2 indicates partner conflict, and Y3 indicates social isolation.

Relation A, conflict with the client's partner is hypothesized to be causing panic episodes which, in turn, affect the client's social isolation. Contrast this to Functional Relation B in which the client's panic episodes are leading to both partner conflict and social isolation. Note how the treatment foci differ based on these assumptions. The treatment focus for "A" would be on partner conflict whereas the treatment focus for "B" would be on panic episodes. For Functional Relation C, the treatment focus might be on panic episodes and/or social isolation.

### ***Strategies of behavioral assessment***

As we noted earlier, a client's adaptive and maladaptive behaviors can be heavily influenced by contextual variables and can vary across response modes, dimensions of responding, and time. These target behaviors are also affected by individual differences—the many unique client characteristics can influence the target behavior. Examples of these unique influences include learning history, biological predispositions, physiological limitations, neurological conditions, social context, gender, race, and ethnicity (Haynes et al., 2018).

Because each client is unique and each target behavior is affected by complex interactions among contextual variables and individual differences, behavioral assessment emphasizes an idiographic approach. An idiographic approach emphasizes the assessment of how a specific client in specific contexts is apt to behave. It uses information derived from empirically supported multimethod, cross-situational, and time-series measurement strategies selected specifically for the client, and focuses on the multidimensional and multimodal nature of a specific target behavior.

### ***Multimethod assessment***

One implication of the aforementioned conceptual foundations is that the use of a multimethod approach in psychological assessment is critically important. By multimethod, we are referring to the use of measures that can capture the various aspects of a client's behavior using different methods of gathering data. In the next section, we provide an overview of these assessment methods, which included the use of direct behavioral observation (e.g., analog behavioral observations and self-monitoring) and self-report (e.g., functional behavioral interviews and questionnaires) methods.

In behavioral assessment, target behaviors can be categorized into three broad classes of response modes: cognitive/verbal responses, affective/physiological responses, and observable behavioral responses. Importantly, different assessment methods are more or less well suited for the measurement of each class of response. When all three responses are congruent, the client's cognitive experience and verbal report of that experience is consistent with both physiological indicators and overt-motor indicators of that experience. Returning to our earlier example, in a congruent state, the client experiencing panic and anxiety would have an awareness of the anxiety and be able to verbally report it during the assessment. At the same

time, the client would show physiological indications of anxiety (e.g., increased heart rate and blood pressure) and overt-motor indications of anxiety (e.g., facial expression and shallow and rapid breathing). In this specific state of congruence, interviewing the client about his or her panic and anxiety would provide a valid index of the client's affective/physiological and overt-motor responses.

However, target behaviors are often characterized by *response desynchrony*. There are six ways that cognitive/verbal, overt-motor, and affective/physiological responses can be desynchronous. Continuing with our panic example, in one state the client may have overt-motor indications of anxiety but no accompanying cognitive/verbal or affective-physiological experience of anxiety. This is commonly observed in clinical contexts and is analogous to “feigning” anxiety. In a second state, the client has physiological indications of anxiety, but does not have the cognitive experience of anxiety and the anxiety is not being expressed overtly in motor responses. This state could be thought of as a lack of awareness or insight. In a third state, the client has a cognitive experience that he or she identifies as anxiety, but there are no physiological indications of anxiety or behavioral expression of anxiety. This state could be described as something analogous to worry and rumination. In a fourth state, the client is experiencing the affective/physiological symptom of anxiety, and is cognitively aware of them, but the anxiety is not expressed overtly in his or her motor responses. This is oftentimes referred to as “inhibition.” In a fifth state, there are physiological and behavioral manifestations of anxiety without client awareness. This state has sometimes been referred to as poor proprioception or “repression.” Finally, in a sixth state, the client is demonstrating overt-motor responses of anxiety and experiencing it cognitively but without any accompanying physiological indicators of anxiety. This state is sometimes referred to as hypervigilance or catastrophic interpretation of body states. The primary point here is that the reliance on a single method of assessment can fail to provide a valid index of complex target behaviors.

An additional rationale for multimethod measurement is that each measurement method has unique sources of error. As noted above, valid self-report requires accurate awareness of physiological and overt-motor responses. However, when a client is asked to report the frequency, intensity, or duration of a behavior across time and contexts, self-report also requires an ability to accurately recall and quantify variation in these responses. Furthermore, because clients have presuppositions about causality, there is a tendency to bias self-reports in a way that reinforces the client's beliefs about causality or is designed to influence the assessor. For example, it is quite common for clients with migraine headache to believe that stress is an important causal variable. However, prospective and longitudinal measurement of the stress–headache relationship is oftentimes incongruent and weaker than the client's self-report (Chiros & O'Brien, 2011).

This discrepancy between retrospective recall and more objective measurement of the stress–headache relationship can reflect a client's neurological impairment and also be linked to an important source of error in verbal reporting of behavior. Specifically, when the client is asked to recall the history of headache causes, the client may selectively attend to and recall instances of “hits” (instances when a

stressful event preceded headache onset) and disregard “misses” (instances when headaches were not preceded by a stressful event) as well as “false positives” (instances when stress was not followed by a headache). This tendency to bias self-reports in favor of presuppositions and heuristics is a commonly observed phenomenon in clinical assessment and was well-articulated by Keller (1903) in her essay on optimism: “I demand the world be good, and lo it obeys. I proclaim the world good, and facts arrange themselves to prove my proclamation overwhelmingly true” (p. 18).

In the second case, a client may bias reports of problem behavior occurrence or causal variable—problem behavior relationships in order to influence the assessor’s behavior. For example, a spouse may report that his wife’s manic episodes are “out of control” and “unpredictable” in order to gain sole custody of their child. Similarly, a client court-ordered for “anger management” may underreport the intensity and/or frequency of aggression in order to be excused from continued therapy.

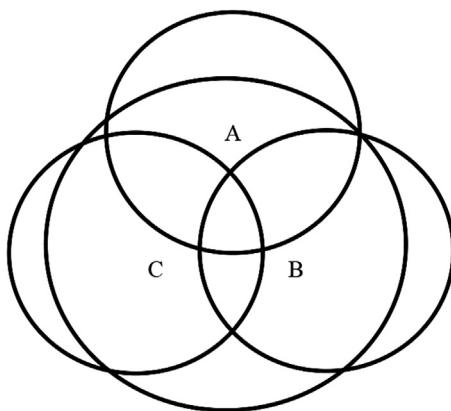
Self-report is not the only assessment method subject to unique sources of error and bias. The reliability, validity, and utility of observational methods can be affected by reactivity to the presence of an observer, the stigma associated with the observed behavior (i.e., some problem behaviors such as sexual responses or verbal conflict hidden from observers), and the degree to which the behavior can be measured using observational methods. Similarly, psychophysiological measures can be affected by sensitivity and intrusiveness of instruments, and strategies used to quantify responses.

Therefore, we strongly recommend that clinicians adopt a multimethod approach in psychological assessment. Specifically, self-reports should be augmented by overt—motor measures of behavior as well as affective/physiological indicators of behavior. This multimethod strategy was difficult to implement in the earlier years of psychological assessment. However, with the wide availability of behavioral and physiological monitoring devices and cell phone applications (see Ecological momentary assessment section), it is quite feasible to use a multimethod approach in behavioral assessment.

## ***Multiple instruments***

Measures from an assessment instrument capture “true variance” in the target behavior and error variance. Additionally, there will be overlap in variance captured among measures of the same target behavior from different instruments. This overlap in variance can be thought of as one indicator of the validity of each instrument (e.g., alternate form validity, predictive validity, and convergent validity). However, each instrument also measures unique aspects of a target behavior that is not shared with other instruments.

Consider Fig. 16.2, which illustrates three different measures of a single target behavior. As is evident in this illustration, there are seven ways that an individual measure or combinations of measures capture variance in the target behavior. Specifically, there are three areas (A, B, C) where each measure alone captures a



**Figure 16.2** Three measures of a single target behavior.

unique source of variance in the target behavior. Then, there are additional areas (A/B, B/C, A/C) where a combination of two measures captures a unique aspect of the target behavior. Finally, there is a seventh area (A/B/C) where a combination of all three measures captures unique variance in the target behavior. It should also be noted that there are additional areas of the target behavior that are not captured by any of the three measures, which is illustrated in the lower part of the larger circle in Fig. 16.2. This unique variance component can contain both error variances as well as “true variances” that are not being measured by this particular set of measures.

Because any measure partially captures unique aspects of a target behavior, and each has idiosyncratic sources of error, no single measure is sufficient. Therefore, based on the conceptual foundations of behavioral assessment pertaining to target behavior complexity and the psychometric limitations of any single instrument, it is important to use multiple instruments to assess a given target behavior. At the minimum, the clinician should strive to use measures that can evaluate each of the main modes of behavioral responses, which are cognitive/verbal responses, affective/physiological responses, and overt—motor responses. Haynes et al. (2019) discuss ways to integrate several different measures to form a composite measure of the target behavior.

### ***Multiinformant assessment***

Because behavior is context dependent, its modes and dimensions will vary according to a myriad of social and environmental factors. Thus, different informants will be apt to have differing information about the modes, magnitude, and temporal characteristics of a client’s target behavior. For example, the coworkers, family members, and friends of a client who experiences panic episodes often interact with and observe the client in very different settings in which panic may be more or less likely to occur. Coworkers can provide valuable and unique information about the

client's behavior during work whereas the client's intimate partner or family member can provide information indicative of the client's behavior in the home and other social settings. Thus, asking only one person to provide a rating of the client's target behavior will fail to identify important sources of variation in the target behavior's magnitude, temporal characteristics, and function.

Another rationale for multiinformant assessment arises from the concerns identified in the prior section, namely presuppositions, heuristics, and biased reporting. For example, there are many studies demonstrating that two or more informants reporting on the same behavior sample for the same person oftentimes have low to moderate levels of agreement (e.g., De Los Reyes et al., 2015). In situations where there is disagreement, the task of the assessor is to separate measurement error from "true variance" in behavior. The question to be addressed is: do the differences among raters reflect true differences in behavior across settings or response modes, systematic bias on the part of one or more informants, or idiosyncratic or unsystematic measurement error (i.e., random uncontrollable errors) associated with each informant?

The often-limited convergence of assessment data among informants highlights the importance of collecting assessment information from the client and other persons with whom the client interacts. Because some of the error in informant reporting is idiosyncratic, the use of multiple informants will allow the clinician to more adequately identify sources of unique informant error from valid information about the form and function of a client's behavior.

### ***Time-series and repeated measurement***

The variation in a client's behavior, in particular it's response modes and dimensions, is continuous and dynamic due to continuously changing contextual variables that affect target behaviors from moment to moment. Thus, a major aspect of behavioral assessment is to design a data collection strategy that will not only capture a person at a particular instance (i.e., a snapshot in time) but also capture the person's moment-to-moment changes in the same behavior. Capturing moment-to-moment variation is important because it helps the clinician to understand the operation of causal variables on a particular target behavior. Recall that one goal of behavioral assessment is to identify the causal variables and causal relations of a client's adaptive or maladaptive behaviors so that they can be targeted during intervention.

There are four common ways to identify the functional relations of a client's behavior. In *event sampling*, the occurrence of a target behavior or causal variable is recorded whenever it occurs during the assessment period. An estimate of frequency or rate of occurrence is then calculated by adding the number of occurrences and dividing them by a relevant time index (e.g., hours, days, or weeks) to get a mean rate of a behavior, such as the average number of "verbal compliments toward spouse per week." Event recording is well suited for target behaviors that have discrete beginning and ending points and occur at a relatively low frequency. This is because high frequency behaviors are difficult to reliably count.

*Duration recording* is used to obtain an estimate of the amount of time that elapses between the onset and end of a target behavior. Duration recording is used when the temporal aspects of a target behavior is most relevant. For example, the amount of time that elapses between going to bed and falling asleep is an important index of one form of sleep onset insomnia. Another example is the length of time between an anger-provoking event and a socially acceptable behavioral response or no response to that event, such as when evaluating the effects of an anger management intervention.

*Interval recording* involves dividing the assessment period into discrete intervals (e.g., seconds, minutes, or hours) and then recording whether the target behavior occurs within a given interval, or not, using dichotomous coding. For example, interval recording can be used to measure off-task behavior of psychiatric inpatients during a 1-h social skills development lesson by dividing the 1-h lesson into 10-min intervals. Either the clinician or a participant observer, such as a psychiatric nurse, then observes the inpatients and each 10-min interval is coded as “on task” or “off task.”

Each of the aforementioned sampling strategies can be used for any type of behavioral assessment instrument. For example, a person could self-monitor the occurrence of suicidal thoughts on a daily basis using event recording, duration recording, or interval recording. Observation instruments and physiological recording instruments can also be designed to quantify target behaviors using event, duration, and interval recording.

### ***Methods of behavioral assessment***

The validity and utility of a clinician’s judgments about a client’s adaptive or maladaptive behaviors and causal factors are affected by his or her ability to measure their dynamic nature across time and settings from multiple sources of information. To increase the ecological validity<sup>4</sup> of the data obtained in clinical assessment, the assessment methods (e.g., direct observation vs self-report questionnaires) and strategies (e.g., event-sampling vs experimental manipulation) used to inform the clinical judgments must accurately capture a client’s multiple behavior problems and causal variables, their functional relations, their changes over time, and across contexts.

In this section, we discuss a diverse set of methods for behavioral assessment with adults in clinical settings, including their conceptual bases, the inferences derived from their use, their limitations, and their clinical utility. Table 16.4 provides a summary of these methods. Recall that we emphasized the use of multiple methods with multiple informants across multiple contexts to improve the validity and utility of the data obtained and of the clinician’s judgments. We also emphasized the use of an idiographic functional approach to assessment. That is, the

<sup>4</sup> *Ecological validity* is the degree to which data obtained during an assessment is representative of, or generalizable to, data that would be derived from the same targets in the environment of primary interest.

**Table 16.4** Summary of the methods of behavioral assessment and their assets and limitations

Method	Example	Assets	Limitations
Naturalistic behavioral Observation	Four 20-min observations of clinic staff–patient interactions on a psychiatric unit	<ul style="list-style-type: none"><li>• Data acquired in the client's natural environment</li><li>• Identification of behavior–environment interactions and functional relations</li><li>• High ecological validity, quantification</li><li>• Amenable to time-series assessment strategies</li></ul>	<ul style="list-style-type: none"><li>• Costly and time consuming</li><li>• Susceptible to assessment reactivity, such as social desirability, when client is aware of observation</li></ul>
Analog behavioral Observation	Clinic-based observation of a couple's attributions while discussing a problem in their relationship	<ul style="list-style-type: none"><li>• Direct observation of behavior in specific situations and conditions</li><li>• Suitable for observation of low-frequency behaviors, such as anger outbursts</li><li>• Cost-effective compared to naturalistic behavioral observation for some behaviors</li><li>• Amenable to time-series assessment strategies and experimental functional analysis (e.g., ABAB designs)</li></ul>	<ul style="list-style-type: none"><li>• Data can have less ecological validity than naturalistic behavioral observations</li><li>• Susceptible to assessment reactivity, such as social desirability</li></ul>
Self-monitoring	Client's daily monitoring of his or her smoking behavior and mood across social settings	<ul style="list-style-type: none"><li>• Data acquired in the client's natural environment</li><li>• Real-time data collection</li><li>• Cost-effective compared to naturalistic and analog behavioral observations</li><li>• Hand-held mobile phones/computers can be used</li><li>• Amenable to time-sampling assessment strategies</li></ul>	<ul style="list-style-type: none"><li>• Problems with adherence to self-monitoring procedures can occur</li><li>• Less useful for client's with major cognitive limitations, such as dementia</li><li>• High reactive effects</li></ul>

Functional Behavioral Interview	Structured interview to identify antecedents and consequences of a client's panic episodes	<ul style="list-style-type: none"> <li>Identification of specific modes and dimensions of behavior</li> <li>Identification of functional relations</li> <li>Cost-effective and time-efficient compared to direct behavioral observational methods</li> <li>Covert and overt behaviors can be measured</li> <li>Measurement of multiple dimensions and response modes of behavior</li> <li>Identification of functional relations, but depends on the degree to which the questionnaire captures the context</li> <li>Low-cost and easy to administer and score</li> </ul>	<ul style="list-style-type: none"> <li>Susceptible to self-reporting biases and errors, such as recall and memory problems</li> <li>Susceptible to interviewer biases, such as skill level</li> </ul>
Functional behavioral Questionnaires	Questionnaire about situations that trigger aggressive or self-harming behaviors	<ul style="list-style-type: none"> <li>Susceptible to self-reporting biases and errors, such as recall problems</li> <li>Predictive validity of measures can vary across instruments</li> </ul>	
Clinic-based Psychophysiological measures	Monitoring heart rate during presentation of anxiety-related stimuli	<ul style="list-style-type: none"> <li>Measurement of multiple dimensions of behavior</li> <li>Amenable to within-subject interrupted time-series designs</li> <li>Amenable to time-sampling and sequential analysis assessment strategies</li> </ul>	<ul style="list-style-type: none"> <li>Cost associated with training of assessors and assessment equipment/devices</li> <li>Ecological validity issues</li> </ul>
Ambulatory biosensors	Ambulatory monitoring heart rate and work-related stressors	<ul style="list-style-type: none"> <li>Data acquired in the client's natural environment</li> <li>Measurement of multiple dimensions of behavior</li> <li>Appropriate for natural and analog settings and self-monitoring</li> <li>High ecological validity</li> <li>Amenable to time-sampling assessment strategies</li> </ul>	<ul style="list-style-type: none"> <li>Cost associated with ambulatory devices and analysis of data</li> </ul>

methods chosen should match the assessment goals and the characteristics of the events being measured for a particular client's behavior problems. For a more detailed discussion on the behavioral assessment methods presented in this section, see [Haynes, O'Brien, et al. \(2011\)](#).

### *Behavioral observation*

Behavioral observation is an important assessment method that can be used to collect data in the client's natural settings where their problem behaviors are most prevalent or in a clinical setting under contrived conditions designed to approximate important aspects of the natural setting. For clinical applications of behavioral observation, we emphasize the quantitative assessment of behaviors based on systematic observation. That is, the systematic recording of observable behaviors using carefully designed procedures to collect reliable and valid data on the behavior and functional relations associated with the behavior ([Bakeman & Haynes, 2015](#); [Haynes, O'Brien, et al., 2011](#)). For example, a clinician may record verbal interactions of a married couple as they discuss a distressing financial issue in a clinic setting and code the frequency and intensity of the negative attributions they express. A clinician might observe an older client living in a care home interacting with members of the nursing staff and other residents to identify specific events (e.g., request by staff members vs social interactions with other residents) that precipitate the client's anger outbursts and differences in their frequency and duration across these contexts.

Behavioral observation methods can address some of the concerns, such as retrospective recall and reporting biases, associated with self-report assessment methods. It is useful for adult clients who have difficulty with speech or cognitive impairments because they cannot provide reliable verbal responses to interview questions, questionnaire items, or other measures requiring self-report. Finally, they are well suited for gathering data on the experimental manipulation of behavior (e.g., ABAB single-subject study) to identify the direction and strength of the relationship between a hypothesized causal variable and a client's behavior problems.

In conducting a behavioral observation, the clinician needs to make several decisions about setting (where the observations should be conducted), coding methods (how to record and quantify observations), sampling strategies, and participants (who should be observed). Essentially, the clinician needs to address several questions. What behavioral problems and putative causal variables or associated events are to be coded? What dimensions of behavior (e.g., frequency, intensity, duration, and onset latency) and characteristics of these variables or events (e.g., their rate of occurrence and topography) are to be coded? In what settings and time of day are these behaviors, causal variables, or events most likely to occur? What time frame (e.g., every 10 min for 1 h) to use for behavior observations? Aside from the client, who else should be the target of observation and coding in the chosen setting? Who will do the observation and coding? The answers to these questions help determine an observation strategy that can capture the target behavior, its putative causal variables, and their functional relations (e.g., response contingencies or triggering events).

Ultimately, the clinician must apply a numerical value (e.g., score) to his or her observations of the relevant dimensions of the target behavior problem and its causal variables and the occurrence of environmental events to quantify these variables and to determine their functional relations. A numerical value is also necessary to monitor the degree of change across settings and over time and change in response to an intervention. At the same time, qualitative data can also be recorded to capture other characteristics of a behavior or causal variable that are not easily quantifiable, such as the form or expression of behavior or unexpected functional relations.

Although it is beyond the scope of this chapter, there are several data-analytic techniques that can be applied to observational data. They include intuitive data analysis (i.e., visual inspection of data using a time-course graph and subjective estimations) and statistical analyses, such as conventional statistical tests, conditional probability analysis, or time-series analysis (e.g., Shadish, Cook, & Campbell, 2002).

### ***Naturalistic behavioral observation***

Behavioral observation in a client's natural environment is highly flexible in regards to the subjects and behaviors targeted for observation and the settings in which the observation takes place, such as in the home, workplace, residential center, or community setting. The observation can be done by the clinician, such as when the clinician codes the frequency of a psychiatric patient's positive interactions with staff members and the consequences on an inpatient unit. It can also be done by someone from the client's natural environment as a participant observer, such as when a nurse codes the intensity of a psychiatric patient's verbal aggression or positive social behavior on an inpatient unit and the events that precede and follow them. The use of participant observers can reduce the reactive effects of the assessment (e.g., when the client alters his or her behavior when being observed) but introduces the possibility of observer bias.

Carefully structured naturalistic observation allows the assessor to capture the situation specificity of a behavior problem. For example, with such data a clinician can evaluate the conditional probability of a client's panic episodes occurring in familiar social settings versus novel social settings. Observation can also permit measurement of highly specific events (e.g., number of people present in the environment when a panic episode occurs) associated with a behavior as well as the dimensions (e.g., onset latency and intensity of panic episode) and functions (e.g., escaping from an anxiety-provoking situation) of that behavior. Thus, behavioral observation can capture quantitative time-series data with high ecological validity and help identify functional relations that are important to a case formulation.

A range of sampling strategies can be used with naturalistic behavioral observation. A *subject-sampling strategy* can be used in which a subsample of subjects is selected from a larger group, such as when several psychiatric patients and staff members from a psychiatric inpatient unit are selected for observation. A *behavior-sampling strategy* can be used in which a subsample of behaviors is selected from a

larger array of behaviors, such as when several forms of aggressive behavior (e.g., positive social behaviors and verbal aggression) are selected for observation of a psychiatric inpatient. A *time-sampling strategy* can be used in which short-periods of observations (e.g., 30 min) are selected over 8 h in a particular day and further subdivided into shorter time segments (e.g., 15 s). Multiple sampling strategies can be used concurrently, such as when rating a patient's positive approaches to staff members (behavior-sampling) every 10 min during a 1-h recreational break (time-sampling) in an inpatient psychiatric setting.

Naturalistic behavioral observation is a powerful method for collecting specific data with good ecological validity. However, it can be difficult or costly to use, particularly for low-frequency behaviors, illegal behaviors (e.g., drug use), socially sensitive behaviors (e.g., domestic violence), and clandestine behaviors (e.g., sexual behaviors). As noted earlier, observational data can also be affected by reactivity (Dishion & Granic, 2004). The observer can also introduce error into the data obtained because of lack of attention, fatigue, or insufficient training. These limitations can be minimized by the careful construction of the behavior codes and scales and the careful training of observers in their use. An index of the reliability of the data collected can be obtained by using multiple observers of the same behavior in the same environment using the same coding scheme and time-sampling strategy.

Participant observers can be less costly for naturalistic behavioral observations and they can diminish the reactive effects of observation. However, they can introduce other biases into the observation, such as when a nurse's prior adverse experiences with a psychiatric patient influences how he or she might rate or label a specific behavior of the patient. They might also have competing obligations in the observation environment, such as when the nurse is observing an inpatient while also being responsible for managing other inpatients. Thus, they are often limited in the number of events they can accurately monitor. For more discussion on the use of participant observers, see LePage and Mogge (2001) and Haynes, O'Brien, et al. (2011).

### ***Analog behavioral observation***

The observation of a behavior in a contrived setting that simulates real-world conditions is referred to as an analog behavioral observation. It is designed to increase the likelihood that target behaviors and relevant functional relations can be captured. For example, a clinician might ask a married couple to discuss a persistent concern about how affection between them is expressed and directly observe the rate of compliments, negative attributions, and interruptions made during the discussion (Snyder et al., 2016).

Several classes of analog behavioral observation methods that can be used, including role-plays, behavioral approach or avoidance test, contrived situation test, experimental functional analysis, think aloud procedure, response generation tests, and enactment analogs (e.g., peer, family, and marital interaction tasks). Several analog behavioral observation protocols have been developed that can be adapted for idiographic clinical applications. The same sampling strategies used for

naturalistic observation (e.g., behavior sampling, subject sampling) can be used in analog observation. As with naturalistic observations, behaviors must be precisely defined, the setting and instructions must be carefully structured, and the observers must be carefully trained. Some examples of these protocols are the *Social Skill Behavioral Assessment System* that uses a role-play test of social skill (Caballo & Buela, 1988) and the *Rapid Marital Interaction Coding System* that uses an enactment analog (Heyman, Weiss, & Eddy, 1995).

Analog behavioral observation is a less costly alternative method to naturalistic behavioral observation and more well suited for low-frequency behaviors. Like naturalistic behavioral observation, data derived from an analog observation can be susceptible to assessment reactivity. For example, a married couple's negative verbal responses to a problematic situation in the presence of the clinician can be less frequent and intense than it would otherwise be in natural setting. Although the rates or intensity of behavior might be affected in analog conditions, important functional relations may still be captured, such as identifying variables that reliably provoke a response (e.g., eye rolling during conversation provokes verbal aggression from a partner).

### ***Self-monitoring***

Self-monitoring is a low-cost alternative to direct behavioral observation methods. It is an assessment method designed to collect a client's self-report data in real-world settings on a wide-range of behaviors. In self-monitoring, the client monitors and records his or her own thoughts, emotions, and overt behaviors along with contemporaneous contextual variables. It is a powerful assessment method for tracking behaviors and changes in behaviors across contexts and time. The behaviors most amendable for self-monitoring are those easily recognizable by the client. For example, a client could record his or her panic episodes, mood, exercise, sleep quality, diet, or mood-altering events across time and settings. Thus, it can be useful for the identification of functional relations, such as the identification of contexts, antecedent events, or consequences that may be contributing to the maintenance of problem behaviors.

As with behavioral observation methods, several issues must be addressed when a clinician designs a self-monitoring system for a client. What behaviors, modes of behaviors, and dimensions of behavior will the client monitor? In what settings will the client self-monitor the behavior? How often will the client monitor the behavior and for how long? How will the client record the behavior and its associated events? What are the factors that can help to facilitate or impede data collection (e.g., competing work demands or degree of cooperation from others in the client's environment)? As far as answering the "how often question," the collection of self-monitoring data can be done either randomly or at fixed times throughout the day (e.g., hourly, after meals, or before bed) or by using event- or experience-sampling in which a client records each occurrence of an event (e.g., depressing stimuli or alcohol use) each time it occurs.

The applicability and utility of using self-monitoring as a clinical assessment method are affected by several factors. A client's biases in reporting and ability to understand and cooperate with self-monitoring procedures can be a source of error. The degree of complexity of the self-monitoring task (e.g., how many behaviors and events must be simultaneously monitored) and how the assessor might use the data (e.g., for determining parental fitness for child custody) can also be sources of error. The client's social environment might pose challenges to accurately self-monitor, such as when the presence of coworkers impedes a client from recording her mood and interpersonal events at the signaled time due to embarrassment. Most of these sources of error can be minimized by providing detailed and simple instructions, carefully designing the data collection procedures, practicing the procedures with the client in advance, and providing follow-up contacts. For examples of self-monitoring, see [Carney et al. \(2012\)](#) for sleep disturbances using a sleep diary; [Faurholt-Jepsen, Munkholm, Frost, Bardram, and Vedel Kessing \(2016\)](#) for mood using information electronic monitoring tools; and [Yacono Freeman and Gil \(2004\)](#) for binge eating using diaries.

Another important source of error associated with using self-monitoring assessment methods is reactivity. Specifically, the act of self-monitoring can lead to changes in the behavior being monitored. Self-monitoring reactivity often leads to changes that are congruent with therapy goals, such as when a client decreases the rate of his cigarette smoking as a result of self-monitoring that behavior ([Moos, 2008](#)). Because of this positive reactive effect, self-monitoring is often used as an intervention strategy (e.g., [Runyan & Steinke, 2013](#)).

Although somewhat costly, self-monitoring data can be made easier to collect and analyze with the use of mobile electronic devices, such as smart phones and other hand-held computers (see the following section). Such mobile electronic devices increase ecological validity of self-monitoring measures and facilitate the collection of specific, real-time data and real-world events. They also facilitate time-sampling strategies, such as random or predesignated sampling, because they can be programmed to signal a client to record behaviors, settings, and antecedent and consequent events. They can also help to reduce some of the sources of error typically associated with traditional paper-and-pencil recording techniques. For more discussion on the use of mobile electronic devices and applications for self-monitoring, see the following section and [Faurholt-Jepsen et al. \(2016\)](#), [Piasecki, Hufford, Solhan, and Trull \(2007\)](#), and [Tsanas et al. \(2016\)](#).

### *Ecological momentary assessment*

EMA is a variation of self-monitoring that is designed to capture within-subject variability and functional relations across time ([Shiffman, Stone, & Hufford, 2008](#)). Its important features are: (1) data are collected by the client in real-world settings as he or she goes about daily life, (2) the client's current state and associated events and contexts are the focus of measurement, (3) data collection focuses on a specific

moment in time, time period, or event, and (4) the client completes multiple measures of the target behavior and events over a specified time period to capture variation across time and contexts and functional relations associated with the target behaviors.

EMA is often used with electronic mobile devices. For example, a client with nicotine dependence can self-monitor the environmental events (e.g., setting, activity, and presence of other smokers) and emotional states that covary with her tobacco use by completing behavioral questionnaires on an electronic diary that randomly signals her to complete four to five times a day over a week. Smartphone applications for EMA are available that allow the EMA to be tailored to the unique behavior problems and assessment goals of the client (e.g., [Conner, 2015](#)). Although EMA is more costly and complicated than other self-monitoring strategies, the use of smartphone applications makes the collection of hundreds of data points easier to manage and analyze.

### *Functional behavioral interviews and questionnaires*

The self-report methods of functional behavioral interviews and questionnaires are the most commonly used methods in behavioral assessment, especially in clinical settings, because they are less costly and easier to administer or conduct than direct behavioral observations and self-monitoring methods. They differ from nonbehavioral types of interviews and queries in their level of specificity for measuring behaviors versus measuring an aggregated cluster of symptoms. For example, a depression questionnaire that assesses symptoms of depression (e.g., negative affect, sleep disturbances, and anhedonia) and their frequency is useful for screening to determine referral for further assessment, or for diagnostic purposes, but it is less useful for identifying the antecedents and consequences of a specific symptom or behavior problem and their conditional probability across settings. Functional behavioral queries go beyond capturing the dimensions of a behavior (e.g., frequency and intensity) to also capturing the contexts, antecedent stimuli, consequences, and/or mediators and moderators that explain variance in the dimensions of a behavior problem: They are designed to *explain* behavior in addition to describing it.

Some examples of functional behavioral questionnaires that capture the dimensions, response modes, and functional relations of a behavior problem are: The Motivation Assessment Scale designed for persons with self-injurious behaviors, but also useful for persons with disruptive, aggressive, or stereotypic behavior problems ([Durand & Crimmins, 1992](#)). The Alcohol Expectancy Questionnaire (adult version) designed to assess anticipated experiences associated with a person's alcohol use ([Brown, Christiansen, & Goldman, 1987](#)). The Revised Conflict Tactics Scale (CTS2) designed to assess the prevalence, frequency, and severity of 39 specific partner conflict behaviors across five categories: "negotiation," "psychological aggression," "physical assault," "sexual coercion," and "injury" ([Straus, Hamby, Boney-McCoy, & Sugarman, 1996](#)). For more discussion on behavioral questionnaires, see [Fernández-Ballesteros \(2004\)](#).

An array of techniques and strategies can be used that include open-ended and closed-ended questions, prompts to elicit information, and observation of paralinguistic behaviors, such as posture, head nods, and facial gestures. Strategies for clarification and elaboration (e.g., paraphrasing and reflections) and providing positive feedback (e.g., highlighting a client's strengths and assets) to establish and maintain a positive and collaborative assessment experience can also be used. An interview strategy to help improve the accuracy of retrospective data collected is the *timeline follow-back* interview procedure (Sobell & Sobell, 1992). It is a semi-structured interview using calendars and memory anchors (e.g., birthdays, anniversaries, and holidays) to construct a daily behavior chart during a specified period of time).

Some examples of queries to specify behaviors and identify functional relations are: "In what situations are you most likely to start drinking?" "How does your wife respond when you start drinking?" "Can you explain to me what your binge drinking is like for you?" "How does it make you feel?" "How many times in the past month have you experienced mood changes that led to your drinking?" "How long did your drinking binge last?" On a scale from one to ten, with ten being the most important, how disruptive has your drinking been for you in regards to your work performance?"

Functional behavioral interviews and questionnaires can be used in conjunction with each other and adapted to fit a wide-range of behaviors and assessment goals and settings. They are particularly useful for obtaining information from multiple informants (e.g., spouses, family members, and psychiatric care staff); on a wide-range of behaviors across different settings; on low-frequency behaviors, such as migraine headaches or panic episodes; on covert or less observable behaviors, such as a client's thoughts, mood, or emotions; on historical events, such as the client's history of trauma and abuse; and on socially sensitive behaviors, such as a client's sexual behaviors and partner violence.

Several factors can affect the validity of data collected from these self-report methods, including biases of the interviewer, the client, or informant; problems in recalling information, or dissimulation. An interviewer might neglect to ask certain questions or overlook or overly attend to certain information provided by the interviewee because of some bias toward the client or because of skill level. A client's or informant's willingness to respond to questions and their ability to remember information or respond honestly can affect the information obtained. Finally, the data obtained from these self-report methods can be affected by certain characteristics of the client or informant, such as their cognitive and language abilities and the severity of their behavior problem. The questionnaires used should be selected based on their validity indices in regards to relevant individual characteristics of the person being assessed, such age, sex, ethnicity, sexual orientation, religious affiliation, and educational level (see Haynes et al., 2018).

### *Psychophysiological assessment*

Psychophysiological assessment methods are often used in behavioral assessment because many behavior problems are associated with important physiological variables, such as when a client misinterprets her increased heart rate as a "heart attack"

or when a client's systolic blood pressure is affected by work-related stressors. They involve the measurement of physiological events, such as blood pressure, heart rate, or cortisol and blood glucose levels, and the variables (e.g., thoughts, emotions, and social interactions) that affect the level, reactivity, variability, or conditional probability of these events. Physiological assessment can occur in analog settings, such as the recording of heart rate in response to a stressor presented in a clinic setting, or in natural settings, such as self-monitoring of blood pressure in response to real-world stressors using biosensors and EMA. These methods are also amenable to multimethod time-series assessment strategies.

[Bandelow et al. \(2000\)](#) study exemplifies the use of psychophysiological assessment. They frequently sampled the salivary cortisol levels of 25 patients with panic disorder in order to identify the neuroendocrine patterns associated with the onset, duration, and intensity of their panic episodes, as measured by several self-report questionnaires. The participants were asked to take a saliva sample at the start of a panic attack and every 5 min for an additional 10 samples (i.e., 180 min in total) and to promptly complete the questionnaires after each sample. They found no significant association between the relative cortisol elevation during the panic episode and the self-reported severity of the episode. This example also illustrates one of our previous points that no single measure is sufficient to capture the target behavior because each may capture unique aspects of a target behavior and each have idiosyncratic sources of error.

Psychophysiological assessments are powerful methods of examining the covariances among physiological, behavioral, and environmental events. They are less susceptible to the biases and recall errors and can provide more accurate and sensitive-to-change measures that are less affected by errors associated with self-report assessment methods. They are particularly useful when a clinician suspects that social desirability or other biases might be affecting a client's self-report; when assessing clients who cannot provide valid self-report data; or when a client is unable to recognize the effects a stimulus is having on his or her behavior.

Psychophysiological assessment is often the most costly of all the assessment methods we described in terms of instrumentation and data analysis, and it can be complicated to use because of the sophisticated technology sometimes involved and the complexity of the data obtained. In deciding whether or not to use psychophysiological assessment methods, the clinician must first consider if psychophysiological responses are an important element of the client's behavior problem. If so, the clinician needs to determine if the psychophysiological data will add to the validity of the data obtained from other methods and yield unique information to aid in a functional analysis or estimating treatment effects (i.e., will it have incremental clinical utility). For more on psychophysiological measurements, see [Cacioppo, Tassinary, and Berntson \(2017\)](#) and [Zisner and Beauchaine \(2015\)](#).

### *Ambulatory biosensor assessment*

Ambulatory biosensor assessment is a type of psychophysiological assessment with a diverse set of strategies to obtain minimally disruptive measures of a client's

motor and physiological responses in the real-world setting (Haynes & Yoshioka, 2007). There are many clinically important physiological and motor responses, such as a client's blood glucose levels, cortisol changes, cardiovascular reactivity, peripheral blood flow, respiration, skin conductance, muscle tension, and physical activity and movement that can be measured using ambulatory biosensors (see Cacioppo, et al., 2017). Ambulatory biosensors can be used with different self-report assessment strategies, such as EMA to examine functional relations between physiological and other responses and contexts (see Giesbrecht, Campbell, Letourneau, Kooistra, & Kaplan, 2012).

For example, the covariance between heart rate variability and airflow obstruction during periods of negative mood and physical activity of a patient with asthma might be examined using self-monitoring strategies in conjunction with ambulatory biosensors (Campbell et al., 2006). The patient's cardiac interbeat intervals could be measured by an electronic ambulatory device and her peak expiratory flow rate measured by an easy-to-use portable peak flow meter. These physiological data could be collected hourly during the day while physical activities (e.g., specific activity, location, and posture), mood (e.g., on a 5-point Likert scale to capture intensity of specific mood states), and self-efficacy (e.g., on a 5-point Likert scale to capture degree of confidence) are concurrently measured. Thus, the use of ambulatory biosensors in behavioral assessment allows for the self-monitoring of specific behaviors using time-sampling techniques, rating scales to capture the dimensions of a behavior, and questionnaires of covert behaviors, such as self-efficacy.

### ***Integrating multiple measures in clinical assessment into a functional analysis***

As is evident in the preceding sections of this chapter, there are many sources of information and many decisions that need to be made during clinical assessment. An important area of decision-making pertains to strategies that can be used to summarize and integrate the complex assessment data. This process of summarizing and integrating assessment information is generically referred to as case formulation or conceptualization. Because the behavioral paradigm emphasizes functional relations, we use the term *functional analysis*.<sup>5</sup> As defined by Haynes and O'Brien (2015) and described in greater detail in Haynes et al. (2019), the functional analysis is "the identification of important, controllable, causal, and noncausal functional relations applicable to specified behaviors for an individual." This definition is consistent with the conceptual foundations and assessment strategies and methods outlined in previous sections of this chapter and in Tables 16.1–16.4: A focus on important functional relations, multiple response modes, in specific contexts, the dynamic nature of behavior, and the use of multiple sources of information.

<sup>5</sup> As we discuss below, the visual-diagrammatic presentation of a functional analysis is referred to as a Functional Analytic Clinical Case Diagram (FACCD). There are multiple definitions and strategies associated with behavioral clinical case formulations and the functional analysis, as outlined in Haynes et al. (2019) and Haynes and O'Brien (2015).

The functional analysis includes three major components: (1) target behaviors (they are the focus of the assessment; they can be problem behaviors to be reduced or goal behaviors to be developed and increased), (2) causal variables, and (3) the functional relations among them. Therefore, one of the critically important steps in constructing a functional analysis is to generate operational definitions of target behaviors and causal variables. In this section, the major dimensions of target behavior and causal variable definitions are reviewed. We will then discuss how to assess causal relations. Finally, we will present a step-by-step strategy for synthesizing the assessment information and presenting it in graphic form—the *functional analytic clinical case diagram* (FACCD).

### ***Operationally defining target behaviors and causal variables***

*Operational definitions of target behaviors.* Behavior can be operationalized along an infinite number of aspects and dimensions. For therapists constructing functional analyses, the dimensions that are most often relevant are: (1) response mode (which can be partitioned into cognitive-verbal, affective-physiological, and overt-motor response systems), (2) response magnitude (such as the peak level of responding or degree of change from a baseline or a comparison condition), and (3) temporal characteristics of response (such as rate and duration). Other dimensions of responding can be relevant in specialized applications of assessment. For example, in psychophysiology, response variability (e.g., heart rate variability, respiratory sinus arrhythmia) and response onset time (e.g., P300 EEG waveform) are important aspects of responding. Identifying and measuring the most relevant dimensions is important because they can be affected by different causal variables and relations. For example, consider how one set of variables can affect the onset of domestic conflict while another affects its intensity and duration.

*Operational definitions of contextual variables.* Contextual variables can also be operationally defined in many ways, but can be broadly partitioned into external/situational, within-person, and historical (e.g., recent history of negative life events). Situational contexts can be further divided into “behavior settings” that can be thought of as built or mentally constructed locations (e.g., relational frames) containing a unique set of stimuli that tend to evoke reliable patterns of behavior (e.g., classroom, bedroom, living room, hospital room, etc.). Additionally, a classroom contains many milieu variables, such as lighting, noise levels, temperature, and seating arrangement as well as interpersonal variables (e.g., number of students, characteristics of students) that can affect target behaviors. Behavior settings, in turn, can be subdivided into (1) milieu variables (e.g., lighting, temperature, and humidity) and (2) interpersonal variables. For example, in a classroom recurrent patterns of behavior can be observed that may be aligned with class scheduling and daily topics. From hundreds of research articles, we also know that behavior can vary as a function of recent and past events, such as negative social interactions, traumatic events, or domestic disputes.

Within-person causal variables are “internal contexts” that are associated with target behaviors. Like target behaviors, within-person causal variables can be

divided into mode, magnitude, and temporal characteristics. Importantly, within-person causal variables may account for variance in a target behavior that is distinct from the influences of setting characteristics, milieu variables, and interpersonal variables.

The important point here is that operationalization and specification are important in clinical case formulation because, as we noted earlier in this chapter, causal relations for behavior problems and goal attainment are conditional. That is, causal relations can be different for the onset of a behavior problem than for its intensity or duration. Similarly, some aspects of a particular environmental context can more strongly affect behavior than can others. For example, in a social situation, the number of persons present, the demands for social interaction, or number of strangers can differentially affect a client's anxiety symptoms. Importantly, these permutations of causal influence can vary across persons with the same behavior problem or goal, highlighting *the idiographic nature of the functional analysis*.

### ***Identifying and evaluating causal relations***

After specifying and generating operational definitions of target behaviors and causal variables, the clinician must identify the strength and direction of the functional relations among them. There are many ways that target behaviors and contextual variables can interact, and it is unlikely that a clinician can reasonably assess all of the possible permutations of interactions and then integrate them into a functional analysis. Thus, science-based a priori causal presuppositions must be made. Causal presuppositions used by clinicians to reduce the complexity of assessment data arise from theory, personal experience, and research on causal relations and individual differences.

In the behavioral assessment literature, the conceptual and science-based foundations described earlier in this chapter form the theoretical basis of causal presuppositions. Additionally, a clinician's training and experience with particular types of clients, target behaviors, and assessment methods will influence the nature of information that is considered important and relevant to the functional analysis. For example, a clinician with psychophysiological training will be apt to include physiological variables into the functional analysis. Finally, clinician individual differences likely influence causality determinations. As with all heuristics, this strategy for identifying causal relations can introduce bias and the clinician must be sensitive to disconfirming evidence. The best defense against bias—a scientific approach to clinical assessment.

*Strategies for evaluating causal relations.* Reliable covariation between a causal variable and some dimension of a target behavior is the essential, albeit preliminary, index of a causal functional relationship. However, covariation alone does not imply causality; the clinician must go one step further and attempt to differentiate causal from noncausal functional relations. Causal and noncausal relations can be differentiated by considering: (1) temporal order—the changes in the causal variable should precede effects on the target behavior, (2) a logical explanation for

the relationship (often based on empirical research and theory), and (3) the exclusion of plausible alternative explanations for the observed relationship.

Three primary assessment strategies can be used to help identify causal functional relationships as well. One approach is the *marker variable strategy*. A marker variable is a measure that is used to index a causal relation. The most commonly used marker in behavioral assessment is a client's report of causality. For example, a client may report that their migraine headaches are triggered by stress or a teacher may report that a student's off-task behavior is for "attention." Other types of marker variables used in behavioral assessment are: scores on questionnaires or rating scales that are used to index a causal relation (e.g., client responses to questionnaire statements indexing a causal event such as "when I am in situation x, my y becomes worse"), observed in-session causal variable–target behavior interactions (e.g., client responses to a therapists statement or action during a session is used as an index of causal relations that occur outside of the session), and physiological measures (e.g., lab-obtained blood glucose levels used to index carbohydrate consumption in the natural environment). Although the marker variable strategy can provide important information about causal relations, few empirically validated marker variables have been developed and used in behavioral assessment. As a result, clinicians typically rely on unvalidated marker variables, such as verbal reports of causal relations obtained during clinical interviews.

A second strategy used to obtain basic information on causal relations is *systematic observation of causal variable–target behavior functional relations*. This is most commonly achieved via client self-monitoring or systematic behavioral observation. Both self-monitoring and direct observation methods can generate valid information about causality. In order to increase the validity of these methods, clients or observers must be carefully trained so that the specified putative causal variable and target behaviors are accurately and reliably recorded in multiple contexts across time as described earlier.

A third method that can be used to identify causal relationships is *experimental manipulation* (sometimes referred to as *experimental functional analysis*). This method involves systematically modifying causal variables while simultaneously observing changes in target behavior. Experimental manipulations can be conducted in naturalistic settings, laboratory settings, and clinical settings.

In summary, marker variables, behavioral observation, and experimental manipulations can be used to gather data on causal functional relations. However, the strength of causal inference associated with each method varies inversely with clinical applicability. That is, experimental manipulations, self-monitoring, and behavioral observation can yield very helpful and valid information about causality, but each method requires a significant investment of time and only a few target behaviors and causal variables can be evaluated at one time. In contrast, while the marker variable strategy has limited or uncertain validity, it is easily administered and can provide information on many potential target behavior-causal variable relationships.

## ***Constructing a functional analysis and functional analytic clinical case diagram***

Once the target behaviors, causal variables, and causal relations have been identified, it is important to synthesize the information in a clinically useful manner in order to inform the design of an intervention and communicate the findings to others. The FACCD is one strategy for accomplishing these two goals. The construction of FACCDs is outlined in detail in [Haynes et al. \(2019\)](#). Here we will provide an abbreviated step-by-step strategy that can guide FACCD construction.

*Step 1: Obtain informed consent.* Informed consent is a collaborative process involving the clinician, the client, and other relevant persons (e.g., family members). The process is designed to help all parties arrive at an agreement about the purposes of the assessment, the relative costs and benefits of the assessment, assessment strategies to be used, and confidentiality.

*Step 2: Evaluate the need for referral and the safety.* After obtaining consent, a clinician should determine whether the nature of the client's problems could be more effectively addressed by another professional. Additionally, in cases where risk for harm is detected, assessment and interventions targeting risk reduction are prioritized over the gathering of additional assessment information.

*Step 3: Identify the target behaviors and treatment goals.* The goal of this step is to identify a wide range of possible target behaviors. Behavioral interviews and self-report inventories are the primary method used to identify behavior problems and treatment goals in this initial stage of the functional analysis. Behavioral observation, such as during a clinical interview, can also help identify behavior problems in some assessment contexts.

*Step 4: Specify target behavior response modes.* Once the client's target behaviors have been identified, specific information about response modes is gathered. In this step, the clinician generates a comprehensive operational definition of each target behavior that will include cognitive, emotional, physiological, and over-motor components.

*Step 5: Specify target behavior dimensions.* The clinician must determine which dimensions are most relevant to the overall functional analysis and FACCD. For some target behaviors, frequency is the most relevant dimension (e.g., number of days absent from work). For other target behaviors, magnitude is most important (e.g., pain intensity). And, in many cases, more than one dimension is relevant.

*Step 6: Estimate target behavior importance.* FACCDs are designed to help the clinician make decisions about intervention foci. A central aspect of this decision is to prioritize target behaviors. Specifically, the clinician aims to determine which target behaviors are most important for client wellbeing and functioning. Thus, a rating of the relative importance of target behaviors is needed.

Relative importance can be estimated in a number of ways. One important consideration is risk of harm to self or others. A second set of considerations is target behavior dimensions: Target behaviors can be prioritized according severity, frequency, and duration. Third, target behaviors can be ranked according to their impact on quality of life. Finally, the clinician must consider the client's subjective rating of importance.

*Step 7: Identify the target behavior effects.* Target behaviors influence other behaviors. For example, the intensity, frequency, or duration of panic attacks can influence interpersonal relationships and social isolation, as noted in our earlier example.

*Step 8: Identify the strength and direction relationships among target behaviors.* Target behaviors can interact in important ways, such as in behavior chains and functional

response classes. They can also have unidirectional or bidirectional causal relations. Identifying the relationships among target behaviors allows the clinician to better estimate the possible impact of an intervention, as illustrated in Fig. 16.1.

*Step 9: Identify causal variables.* Identifying causal variables is one of the most critical steps in the generation of a FACCD because most interventions involve the modification of causal variables in order to exert an effect on target behaviors. For example, if a client's depressed mood were mainly caused by marital distress, then marital therapy would be an appropriate intervention. Alternatively, if a client's depressed mood were mainly caused by social isolation, then social skills training would be an appropriate intervention.

*Step 10: Estimate the modifiability of causal variables.* Causal variables can differ in terms of modifiability. A causal variable that can be modified (e.g., social behaviors) is more relevant to intervention design than unmodifiable or less modifiable causal variables (e.g., childhood abuse, traumatic brain damage).

*Step 11: Estimate the strength and direction of relations among causal variables and target behaviors.* After identifying the important and modifiable causal variables, the clinician will then estimate the strength and direction of relationships among these variables.

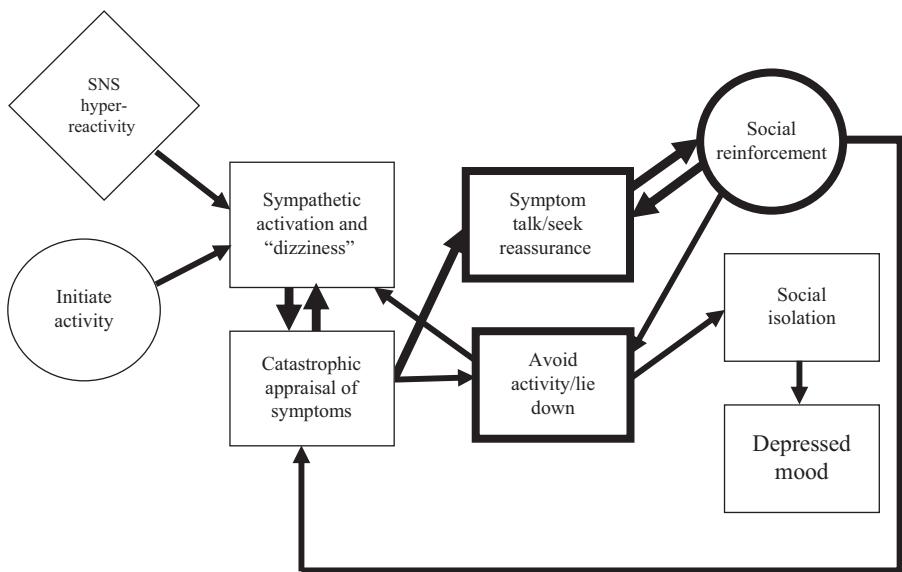
*Step 12: Estimate the strength and direction of causal relations among causal variables.* There can be unidirectional and bidirectional causal relations of varying strength among causal variables. As with target behaviors, these causal relations influence the selection of the best intervention focus. Modifying a causal variable that has multiple or stronger functional relations with other causal variables and target behaviors will result in a larger intervention effect than a causal variable that has fewer or weaker functional relations.

*Step 13: Identify potential moderator and mediator variables.* A moderator variable alters the strength or the direction of a causal relation between two other variables. A moderator variable can increase, decrease, and in some cases reverse the effects of a causal variable. For example, "social support" is a commonly referenced moderator variable that alters the strength of the relations between stressor exposure and health problems. Mediators help "explain" a causal relation.

*Step 14: Generate a FACCD.* The prior steps involve gathering information about the target behaviors, causal variables, and relations among them. In the final step, the information can be synthesized using vector diagrams that are analogous to structural equation modeling. As illustrated in Fig. 16.3, target behaviors and causal variables are differentiated by the shapes in the vector diagram. Importance ratings of these same variables are captured using line thickness of the shapes (low, medium, high importance). The functional relations among variables are depicted by vector lines that are either unidirectional or bidirectional. The strength of relations is depicted by vector line thickness (small, medium, large).

### ***Sam: an example of behavioral assessment strategies, a functional analysis, and FACCD***

An example of a FACCD is presented in Fig. 16.3. The client was a 26-year-old male (note: identifying information has been changed to preserve confidentiality) who was referred for treatment of a psychogenic pseudosyncope. During the initial interview, Sam reported that he was experiencing severe fainting symptoms and dizziness whenever he attempted to walk or move from a reclining position. Sam had been evaluated by dozens of physicians from varied specialties including



**Figure 16.3** A functional analytic case diagram of a client presenting with a diagnosis of psychogenic pseudosyncope. The diagram illustrates unidirectional and bidirectional relations among multiple behavior problems and the influential roles of anxiety avoidance and social reinforcement.

neurology, cardiology, and immunology. None of the evaluations yielded an identifiable organic or medical basis for Sam's symptoms. He was thus referred for a behavioral assessment with accompanying possible diagnoses of "dysautonomia," "postural orthostatic tachycardia," and "conversion disorder."

The behavioral assessment began with an interview in Sam's home (he resided with his parents) because he reported he was unable to travel to the outpatient clinic due to symptoms. It was observed that Sam was able to sit up and talk about his difficulties with great detail and enthusiasm. He reported that his symptoms began shortly after he obtained his undergraduate degree when he began feeling pressure from his parents to obtain a job and look into living independently. He felt the onset of symptoms was insidious and relentless, ultimately causing him to lose all ability to leave the home. He further noted that he was very active with online support groups and that his mother was providing intensive and nurturing daily care. She would also spend evenings with him discussing his symptoms and making sure he was not alone. Sam reported that, despite his mother's care, he found himself experiencing social isolation and depressed moods. He noted that he had given up on the idea that he could live a "normal life."

After the initial interview, Sam was asked to self-monitor dizziness, fainting symptoms, depressed mood, anxiety levels, activity level, and location four times per day. The results of self-monitoring, further interviewing, parent monitoring of activity levels, and a subsequent psychophysiological assessment (we measured

blood pressure, heart rate variability, and heart rate during postural adjustments, such as standing and sitting) were integrated into a FACCD as illustrated in Fig. 16.3.

The FACCD has a number of elements. First, modifiable causal variables are identified by circles, unmodifiable casual variables by diamonds, and target behaviors by rectangular boxes. The degree of importance of target behaviors and modifiability of causal variables is depicted by the thickness of shape outlines with thicker lines indicating a higher degree of importance and modifiability. The direction of relationships among variables is depicted by vector lines. The strength of relationships is depicted by line thickness with thicker lines denoting a stronger relationship.

As is evident in Fig. 16.3 and consistent with the principles of behavioral assessment presented in this chapter, we see in this case: Multiple and complex functional relations (unidirectional, bidirectional, different strengths) among multiple causal variables and multiple behavior problems. Further, the modifiability of causal variables differs. Finally, there are multiple influences on target behaviors.

Despite the complexity of relationships depicted in Fig. 16.3, an evaluation of the causal paths, the weights associated with causal paths, and the modifiability of variables can aid in treatment design. In this particular instance, the FACCD suggests the following sequence. Upon being prompted to engage in some type of physical activity, Sam would experience sympathetic nervous system activation and a sense of dizziness. Upon experiencing these symptoms, Sam generated catastrophic appraisals (e.g., something dreadfully wrong was happening in his body and terrible consequences were imminent). These catastrophic appraisals would then intensify sympathetic activation and the associated symptoms which, in turn, intensified the intensity and believability of his catastrophic appraisals. Sam would thus very quickly find himself experiencing intense sympathetic symptoms and intense catastrophic thoughts. In order to assuage his fear, Sam would seek reassurance from others and withdraw from physical activity and lying down. These two actions were followed by immediate and powerful reductions in sympathetic activation and catastrophic thoughts. As such, they were functioning as important sources of negative reinforcement. Finally, social isolation and depressed mood were an outcome of the restricted range of activities and limited social contact with others.

An intervention that combined acceptance/mindfulness-based behavior therapy, exposure therapy, contingency management, and communication skills training was implemented with the FACCD in mind. The intervention addressed catastrophic appraisals using cognitively focused techniques derived from mindfulness-based behavior therapy (acceptance, mindfulness, and defusion). The exposure therapy targeted reduction of avoidance behaviors and acquisition of adaptive behaviors by having Sam engage in increasing levels of daily age-appropriate and goal-directed activities, such as meal preparation, household chores, driving, engaging in real-world social interactions, and job seeking. The behavioral contracting/contingency management intervention involved meeting with Sam's mother to provide guidance on how to systematically and consistently provide positive reinforcement for adaptive behavior and simultaneously reduce reinforcement for problematic behaviors.

Finally, communication skills training was provided to assist Sam with forming relationships without relying on “symptom talk” as the primary topic of conversation. These skills were initially practiced with his mother because it was observed that both Sam and his mother needed to develop a more functional communication pattern. After practicing with his mother, Sam was then encouraged to practice communication skills in criterion situations, such as the university student union, church social events, and social gatherings with peers.

### ***It makes a difference: the effects of interventions based, or not based, on the functional analysis***

Hurl, Wrightman, Haynes, and Virues-Ortega (2016) compared the relative effectiveness of interventions based on a pre-intervention functional behavioral assessment (FBA) with interventions not based on a pre-intervention FBA (e.g., “best practice” interventions) in 19 single-case studies. Participants were mostly persons with developmental disabilities. Effect sizes, based on a random effects meta-analysis, indicated that FBA-based interventions were associated with significantly larger reductions in problem behaviors compared to non-FBA-based interventions and no intervention. FBA-based and non-FBA-based interventions were associated with significant increases in appropriate behavior relative to no intervention, but the effects were much larger for FBA-based interventions. These results were consistent with several previous literature reviews, as discussed in Hurl et al. (2016) of less-well controlled studies on the effectiveness of basing intervention decisions on the results of a functional analysis.

## **Summary and concluding recommendations**

Based on the conceptual foundations, strategies, and methods of behavioral assessment and the integration and communication of assessment information into a functional analysis, and regardless of assessment paradigm, the clinical assessment of adults will result in information that is more valid and useful for intervention decisions to the degree that the assessment process:

1. *Carefully specifies* a client’s behavior problems and treatment goals on *multiple dimensions* and recognizes that a client may have *many problems and goals that interact* in complex ways.
2. Expands the focus beyond problem identification and diagnosis and strongly attends to the *causal variables and functional relations* associated with behavior problems and goals, including their strength, form, direction, and interactions.
3. Identifies *mediators and moderators* of important causal relations for a client’s behavior problems and goals.
4. Uses measures that have *strong psychometric evidence* that is relevant for the client and the goals and context of assessment.

5. Includes strategies and methods that indicate respect, and are psychometrically appropriate, for the client's *dimensions of individual difference* (e.g., ethnicity, age, and economic status).
6. Includes *multiple assessment instruments* to measure *multiple response modes*, utilizing data from *multiple informants*.
7. Includes strategies that are sensitive to the *dynamic nature* of behavior and causal relations and includes ongoing assessment during intervention and follow-up.
8. Reflects the *context dependent* and *conditional nature* of behavior and its causal relations, particularly focusing on *environmental contexts* and *contemporaneous antecedent and consequent events for behavior*, and the client's *extended social network*.
9. Includes measures and clinical judgment that are *more direct and less inferential* than measures that depend only on retrospective recall, responses to projective stimuli, or tests of personality traits.
10. Results in a valid and useful *clinical case formulation* that integrates data on all important variables and functional relations.
11. Includes strategies that are sensitive to the *dynamic nature* of behavior and causal relations and includes ongoing assessment during intervention and follow-up.

## References

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.
- Bakeman, R., & Haynes, S. N. (2015). Behavioral observation. In R. L. Cautin, & S. O. Lilienfeld (Eds.), *The encyclopedia of clinical psychology*. New York, NY: Wiley-Blackwell.
- Bandelow, B., Wedekind, D., Pauls, J., Broocks, A., Hajak, G., & Ruther, E. (2000). Salivary cortisol in panic attacks. *American Journal of Psychiatry*, 157(3), 454–456. Available from <https://doi.org/10.1176/appi.ajp.157.3.454>.
- Beck, A. T., & Steer, R. A. (1990). *Manual for the beck anxiety inventory*. San Antonio, TX: Psychological Corporation.
- Brown, S. A., Christiansen, B. A., & Goldman, M. S. (1987). The alcohol expectancy questionnaire: An instrument for the assessment of adolescent and adult alcohol expectancies. *Journal of Studies on Alcohol*, 48(5), 483–491. Available from <http://dx.doi.org/10.15288/jsa.1987.48.483>.
- Budd, K., Connell, M., & Clark, J. (2011). *Evaluation of parenting capacity in child protection*. New York, NY: Oxford University Press.
- Caballo, V. E., & Buela, G. (1988). Factor analyzing the college self-expression scale with a Spanish population. *Psychological Reports*, 63, 503–507.
- Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2017). *Handbook of psychophysiology* (4th ed.). Cambridge, MA: Cambridge University Press.
- Campbell, T. S., Lavoie, K. L., Bacon, S. L., Scharf, D., Aboussafy, D., & Ditto, B. (2006). Asthma self-efficacy, high frequency heart rate variability, and airflow obstruction during negative affect in daily life. *International Journal of Psychophysiology*, 62, 109–114. Available from <https://doi.org/10.1016/j.ijpsycho.2006.02.005>.
- Carney, C. E., Buysse, D. J., Ancoli-Israel, S., Edinger, J. D., Krystal, A. D., & Lichstein, K. L. (2012). The consensus sleep diary: Standardizing prospective sleep

- self-monitoring. *Sleep*, 35(2), 287–302. Available from <http://dx.doi.org/10.5665/sleep.1642>.
- Chiros, C., & O'Brien, W. H. (2011). Acceptance, appraisals, and coping in relation to migraine headache: An evaluation of interrelationships using daily diary methods. *Journal of Behavioral Medicine*, 34, 307–320. Available from <https://doi.org/10.1007/s10865-011-9313-0>.
- Conner, T.S. (May 2, 2015). Experience sampling and ecological momentary assessment with mobile phones. Retrieved from <http://www.otago.ac.nz/psychology/otago047475.pdf>.
- De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, 141(4), 858–900. Available from <https://doi.org/10.1037/a0038498>.
- Dishion, T. J., & Granic, I. (2004). Naturalistic observation of relationship processes. In E. M. Heiby, & S. N. Haynes (Eds.), *Comprehensive handbook of psychological assessment: Vol. 3. Behavioral assessment* (pp. 143–161). New York, NY: Wiley.
- Durand, V. M., & Crimmins, D. B. (1992). *The motivation assessment scale (MAS) administration guide*. Topeka, KS: Monaco & Associates.
- Ebner-Priemer, U. W., & Trull, T. (2009). Ambulatory assessment: An innovative and promising approach for clinical psychology. *European Psychologist*, 14(2), 109–119. Available from <http://dx.doi.org/10.1027/1016-9040.14.2.109>.
- Eckert, T. L., & Lovett, B. J. (2013). Principles of behavioral assessment. In D. H. Saklofske, C. R. Reynolds, & V. Schwean (Eds.), *The Oxford handbook of child psychological assessment* (pp. 366–422). New York, NY: Oxford University Press.
- Faurholt-Jepsen, M., Munkholm, K., Frost, M., Bardram, J. E., & Vedel Kessing, L. (2016). Electronic self-monitoring of mood using IT platforms in adult patients with bipolar disorder: A systematic review of the validity and evidence. *BMC Psychiatry*, 16. Available from <https://doi.org/10.1186/s12888-016-0713-0>, Article 7.
- Faurholt-Jepsen, M., Vinberg, M., Frost, M., Debel, S., Christensen, E. M., Bardram, J. E., & Kessing, L. V. (2016). Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. *International Journal of Methods in Psychiatric Research*, 25(4), 309–323.
- Fernández-Ballesteros, R. (2004). Self-report questionnaires. In S. N. Haynes, & E. M. Heiby (Eds.), *Comprehensive handbook of psychological assessment: Vol. 3. Behavioral assessment* (pp. 194–221). Hoboken, NJ: John Wiley & Sons, Inc.
- Fisher, J., O'Donohue, W. W., & Haynes, S. N. (2018). Introduction to principles and foundations of psychological assessment. In P. Sturme (Ed.), *Aggression and violence*. Hoboken, NJ: John Wiley & Sons, Inc.
- Geisinger, K. F. (Ed.), (2013). *APA handbook of testing and assessment in psychology* (Vols. I-III). Washington, DC: American Psychological Association.
- Giesbrecht, G. F., Campbell, T., Letourneau, N., Kooistra, L., & Kaplan, B. (2012). Psychological distress and salivary cortisol within persons during pregnancy. *Psychoneuroendocrinology*, 37, 270–279. Available from <https://doi.org/10.1016/j.psyneuen.2011.06.011>.
- Granacher, R. P. (2015). *Traumatic brain injury: Methods for clinical and forensic neuropsychiatric assessment* (3rd ed.). Boca Raton, FL: CRC Press.
- Guion, R. M. (2011). *Assessment, measurement, and predictions for personnel decisions*. New York, NY: Taylor & Francis.

- Hart, A. J. P., Gresswell, D. M., & Braham, L. G. (2011). Formulation of serious violent offending using multiple sequential functional analysis. In P. Sturmey, & M. McMurran (Eds.), *Forensic case formulation* (pp. 81–106). Hoboken, NJ: John Wiley & Sons.
- Haynes, S. N., Kaholokula, J. K., & Tanaka-Matsumi, J. (2018). Psychometric foundations of psychological assessment with diverse cultures: What are the concepts, methods, and evidence? In W. O'Donohue, & C. Frisby (Eds.), *Cultural competence in applied psychology: Theory, science, practice, and evaluation*. New York, NY: Springer Publishing Co.
- Haynes, S. N., Mumma, G. H., & Pinson, C. (2009). Idiographic assessment: Conceptual and psychometric foundations of individualized behavioral assessment. *Clinical Psychology Review*, 29, 179–191. Available from <https://doi.org/10.1016/j.cpr.2008.12.003>.
- Haynes, S. N., & O'Brien, W. H. (2015). Functional analysis. In R. L. Cautin, & S. O. Lilienfeld (Eds.), *Encyclopedia of clinical psychology*. New York, NY: Wiley-Blackwell.
- Haynes, S. N., O'Brien, W. H., & Kaholokula, J. K. (2011). *Behavioral assessment and case formulation*. Hoboken, NJ: John Wiley & Sons.
- Haynes, S. N., Smith, G., & Hunsley, J. R. (2019). *Scientific foundations of clinical assessment* (2nd ed). New York: Taylor and Francis/Routledge.
- Haynes, S. N., & Yoshioka, D. T. (2007). Clinical assessment applications of ambulatory bio-sensors. *Psychological Assessment*, 19(1), 44–57. Available from <https://doi.org/10.1037/1040-3590.19.1.44>.
- Heilbrun, K., DeMatteo, D., Brooks Holliday, S., & LaDuke, C. (2014). *Forensic mental health assessment: A casebook*. New York, NY: Oxford University Press.
- Hersen, M. (Ed.), (2006). *Clinician's handbook of child behavioral assessment*. San Diego, CA: Elsevier Academic Press.
- Heyman, R. E., Weiss, R. L., & Eddy, J. M. (1995). Marital interaction coding system: Revision and empirical evaluation. *Behaviour Research and Therapy*, 33(6), 737–746. Available from [https://doi.org/10.1016/0005-7967\(95\)00003-G](https://doi.org/10.1016/0005-7967(95)00003-G).
- Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future-remembering the past. *Annual Review in Psychology*, 51, 631–664. Available from <https://doi.org/10.1146/annurev.psych.51.1.631>.
- Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, 12(2), 205–218. Available from <https://doi.org/10.1037/1082-989X.12.2.205>.
- Hurl, K., Wrightman, J., Haynes, S. N., & Virues-Ortega, J. (2016). Does a pre-intervention functional assessment increase intervention effectiveness? A meta-analysis of within-subject interrupted time-series studies. *Clinical Psychology Review*, 47, 71–84. Available from <https://doi.org/10.1016/j.cpr.2016.05.003>.
- Kazdin, A. E. (2001). *Behavior modification in applied settings* (6th ed.). Belmont, CA: Wadsworth/Thomson Learning.
- Keller, H. (1903). *Optimism*. New York, NY: TY Crowell and Co.
- Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology*, 2, 111–133. Available from <https://doi.org/10.1146/annurev.clinpsy.2.022305.095213>.
- Lepage, J. P., & Mogge, N. L. (2001). The behavioral observation system (BOS): A line staff assessment instrument of psychopathology. *Journal of Clinical Psychology*, 57(12), 1435–1444. Available from <https://doi.org/10.1002/jclp.1107>.

- Maddux, J. E., & Winstead, B. A. (Eds.). (2012). *Psychopathology: Contemporary issues, theory, and research* (3rd ed.). Mahwah, NJ: Erlbaum.
- McGrane, J. A. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology*, 6, Article 431. Available from <http://dx.doi.org/10.3389/fpsyg.2015.00431>.
- McLeod, B. D., Jensen-Doss, A., & Ollendick, T. H. (Eds.). (2013). *Diagnostic and behavioral assessment in children and adolescents*. New York, NY: Guilford Press.
- Moos, R. H. (2008). Context and mechanisms of reactivity to assessment and treatment: Comment. *Addiction*, 103(2), 249–250.
- O'Brien, W. H., Keawe'aimoku Kaholokula, J., & Haynes, S. N. (2016). Behavioral assessment and functional analysis. In A. Nezu, & C. Nezu (Eds.), *The oxford handbook of cognitive behavioral therapies* (pp. 44–61). New York: The Oxford University Press.
- Persons, J. B. (2008). *The case formulation approach to cognitive-behavior therapy*. New York, NY: Guilford Press.
- Piasecki, T. M., Hufford, M. R., Solhan, M., & Trull, T. J. (2007). Assessing clients in their natural environments with electronic diaries: Rationale, benefits, limitations, and barriers. *Psychological Assessment*, 19(1), 25–43. Available from <https://doi.org/10.1037/1040-3590.19.1.25>.
- Runyan, J. D., & Steinke, E. G. (2015). Virtues, ecological momentary assessment/intervention and smartphone technology. *Frontiers in Psychology*, 6, Article ID 481.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage Learning.
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32. Available from <https://doi.org/10.1037/0021-843X.115.3.509>.
- Snyder, D. K., Heyman, R. E., Haynes, S. N., Carlson, C. I., & Balderrama-Durbin, C. (2016). Couple and family assessment. In J. C. Norcross, G. R. VandenBos, & D. K. Freedheim (Eds.), *APA handbook of clinical psychology: Vol. 3. Applications and methods*. Washington, DC: American Psychological Association.
- Sobell, L. C., & Sobell, M. B. (1992). Timeline follow-back: A technique for assessing self-reported ethanol consumption. In R. Z. Litten, & J. Allen (Eds.), *Measuring alcohol consumption: Psychosocial and biological methods*. (pp. 41–72). New Jersey: Humana Press.
- Sturmey, P. (2008). *Behavioral case formulation and intervention: A functional analytic approach* (1st ed.). New York, NY: Wiley-Blackwell.
- Straus, M. A., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). The revised conflict tactics scales (CTS2): Development and preliminary psychometric data. *Journal of Family Issues*, 17(3), 283–316. Available from <https://doi.org/10.1177/019251396017003001>.
- Tsanas, A., Saunders, K. E., Bilderbeck, A. C., Palmius, N., Osipov, M., Clifford, G. D., ... De Vos, M. (2016). Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder. *Journal of Affective Disorders*, 205, 225–233. Available from <https://doi.org/10.1016/j.jad.2016.06.065>.
- Yacono Freeman, L. M., & Gil, K. M. (2004). Daily stress, coping, and dietary restraint in binge eating. *International Journal of Eating Disorders*, 36(2), 204–212. Available from <https://doi.org/10.1002/eat.20012>.

Zisner, A., & Beauchaine, T. P. (2015). Physiological methods and developmental psychopathology. In D. Cicchetti (Ed.), *Developmental psychopathology* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

## Further reading

O'Neill, R. E., Albin, R. W., Storey, K., Horner, R. H., & Sprague, J. R. (2015). *Functional assessment and program development for problem behavior: A practical handbook* (3rd ed.). Stamford, CT: Cengage Learning.

## **Part IX**

# **Special Topics and Applications**

# Psychological assessment of the elderly

17

Jeannie Lengenfelder<sup>1</sup>, Karen L. Dahlman<sup>2</sup>, Teresa A. Ashman<sup>3,4</sup> and Richard C. Mohs<sup>5</sup>

<sup>1</sup>Department of Physical Medicine & Rehabilitation, Rutgers, the State University of New Jersey, New Jersey Medical School, Newark, NJ, United States, <sup>2</sup>Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, United States, <sup>3</sup>Private Practice, New York, NY, United States, <sup>4</sup>NYU Langone Medical Center, New York, NY, United States, <sup>5</sup>Global Alzheimer's Platform Foundation, Washington, DC, United States

## Introduction

The proportion of the population over age 65 in the United States is projected to increase to 16% by 2020 (Dept of Economic and Social Affairs, 2010). As the proportion of the population over age 65 increases, the field of geropsychology continues to expand exponentially. Incorporated into geropsychology are issues related to the psychological assessment of the elderly, including the appropriate application and limitations of psychometric tools. The intent of this chapter is to present a context in which to understand the cognitive impact that occurs as a process of aging. The first part will examine the normal aging process. The second part outlines the importance of gathering a complete clinical picture of premorbid and present functioning exclusive of cognitive abilities (e.g., medical conditions, family history, social adaptation, psychiatric history). The third part focuses on cognitive functioning among older patients, including methods of determining dementia. The fourth part highlights the principles of the neuropsychological assessment process for older adults, along with the purposes utilized by the assessment. The fifth part highlights typical differential diagnosis questions, such as Alzheimer's disease (AD) versus depression.

Prevalence rates from a national survey indicate a prevalence of 11.9% for depression among older American (Steffens et al., 2009) Dementia has been reported to occur in 13.9% of the population aged 71 years or older (Plassman et al., 2007). Given the prevalence of both depression and dementia in the geriatric population of the United States, the need for accurate assessment of both conditions is vital (Bowler et al., 1994; Katzman, Lasker, & Berstein, 1988; Lamberty & Bielaukas, 1993). Early diagnosis of dementia is more important than ever, with the introduction of new cognition-enhancing pharmacological agents (Samuels & Davis, 1998).

Current approaches to the psychological assessment of the elderly, like that of any other patient group, have developed over time. Early approaches (Adams, 1986; Adams & Heaton, 1985; Reitan & Davidson, 1974; Reitan & Wolfson, 1993;

Russell, Neuringer, & Goldstein, 1970; Swiercinsky, 1978) have reflected trends in the field of psychology to become more medical model, focusing on such constructs as deficits, and with diagnoses derived by use of actuarial formulas based on testing and/or symptom checklist. Other approaches (Christensen, 1979; Luria, 1966, 1973) have drawn on case study literature and have emphasized careful clinical observation of the patient. This model has early associations with Luria and later advocated by Lezak (1995), Lezak, Howieson, Bigler, and Tranel (2012) who argues that assessment should integrate qualitative behavioral descriptions, examinations of patients' writings and drawings, and attempts to elicit behaviors that reflect brain function as well as quantitative instruments. This approach also involves testing of hypotheses that guide clinical exploration, diagnosis, and formulation of treatment recommendations (Kaplan, 1988). Current methods of assessing older adults have similar goals as those assessing other ages and include utilizing testing and assessment to solve real world problems and discover strengths and weaknesses (Ornum et al., 2008).

Additionally, a thorough evaluation of the older patient must be function-based; that is, it must include assessments of level of functioning. Older patients with medical, cognitive and/or psychological problems also have functional and support issues that strongly affect their quality of life. Areas of concern to the assessor must include basic and instrumental activities of daily living, cognition, mood, psychiatric and medical diagnoses, balance and mobility, sensory intactness, continence, nutritional status, and living arrangements.

Methodological issues are present in studies examining the aging population. Many studies utilize a cross-sectional design to evaluate participants from different age groups. An issue with utilizing cross-sectional design, although easier to carry out a study on, is that the cohorts may have differences due to life experiences, education, work life, and culture (Hedden, 2004; Williams, 1996). Although longitudinal studies are able to evaluate participants over time, issues of attrition, practice effects, or cognitive dysfunction due to pathologic rather than normal aging that emerges could make such work problematic and difficult to generalize. Therefore as in other research, work in assessment of the elderly is still needed.

## **Normal aging**

A key issue in psychological assessment of elderly patients is the need to discriminate between normal age-related intellectual changes and those changes that are clinically significant. The aged are at least as varied a population as teens, college students, or middle-aged individuals. Some will change very little as they age, others a great deal, and still to others the change will be in only a few areas. Therefore it is useful to know what cognitive functions normally decline with age as well as what impairments are common for age-related conditions like AD.

Although some cognitive functions decline as a part of the normal aging process (Wechsler, 1997a, 1997b), the extent and pattern of the decline varies according to both the individual and the type of function being examined. Cognitive abilities that deal with well-rehearsed, overlearned information change very little across the

lifespan. Cognitive abilities such as vocabulary remain preserved as individuals age. In fact, more recent evidence suggests that these abilities may even improve slightly in later years (Salthouse, 2012). Other cognitive functions, including memory, executive abilities, and processing speed tend to decline as individuals age (Harvey & Dahlman, 1998; Salthouse, 2012).

Memory assessments typically focus on declarative or nondeclarative memory abilities. Declarative memory, or explicit memory, is the recall or recognition of facts or events, such as knowing a dog is an animal or the name of your first-grade teacher. Declarative memory has been shown to decline with normal aging (Ronnlund, 2005). Nondeclarative memory, or implicit memory, does not require conscious thought and is often procedural in nature, such as riding a bike or brushing your teeth. Nondeclarative memory is not as susceptible to age-related declines in the same way as declarative memory (Cargin, 2007; Price, 2004).

Executive functioning encompasses varied higher order abilities including planning, reasoning, cognitive flexibility, abstraction, inhibition, and initiation (Lezak et al., 2012).

The considerable individual differences in cognitive changes with aging indicate not only the difference between normal and impaired changes over time, but also differences between normal and successful changes as individuals age. Using the example of normative standards on the Logical Memory subtest from the Wechsler Memory Scale (Wechsler, 1997b), it becomes clear that those individuals who performed at high levels (99th percentile) in their youth on a variety of cognitive domains tend to decline very little throughout their lifespan. Individuals who performed at lower levels (15th percentile) in their youth exhibit not only a decline, but a sharper decline than individuals in the upper percentile scores. The individuals at the top of the distribution consistently outperform those at the lower levels by a progressively greater extent as they become older.

The idea that normal adults who perform at higher baseline levels of intellectual function will exhibit little cognitive decline with age is supported by Rowe and Kahn's (1987, 1997) reports on successful aging. They define successful aging as including three main components: low probability of disease and disease-related disability, high cognitive and physical functioning, and active engagement with life. Continuing engagement with life has two major elements: maintenance of interpersonal relations and productive activities. Membership in a social network is an important determinant of longevity (House, Landis, & Umberson, 1988). Network membership research (Cassel, 1976; Glass, Seeman, Hertzog, Kahn, & Berkman, 1995; Kahn & Byosiere, 1992) has demonstrated that two types of supportive transactions may be prophylactic in aging: socio-emotional and instrumental. Socio-emotional transactions include expressions of respect and affection, while instrumental transactions are comprised of direct giving of services or money.

It is critical in the assessment of elderly individuals to take into account the relative nature of observed deficits; relative, that is, to the patient's own previous levels of functioning. Current functioning, in terms of engagement in life as well as presence/absence of disease and cognitive normalcy, must be viewed against the individual's overall level of previous functioning. Even a clinical interview of the patient

combined with neuropsychological testing may be not be enough to fully assess what the patient may have been like prior to the onset of symptoms (Harwood, Hope, & Jacoby, 1997a; Harwood, Hope, & Jacoby, 1997b; Williams, 1997). For this reason, there is a trend to include caregiver ratings of patients as part of the assessment process. There are many caregiver ratings available for use that cover a variety of abilities such as activities of daily living [Caregiver Assessment of Function and Upset (CAFU), Gitlin et al., 2005; Bristol Activities of Daily Living Scale (BADLS), Bucks et al., 1996; Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE), 1989], behavioral issues (Frequency of Behavior Problems Scale, Neiderehe & Fruge, 1984), and quality of life [DEMQOL, Smith et al., 2007; Alzheimer's Disease-related Quality of Life scale (QoL-AD), Logsdon et al., 1999].

## Clinical assessments

A thorough history is necessary to establish premorbid levels of functioning in all areas of the individual's life. It should include relevant medical, family, social, occupational, educational, cultural, and medication history, as well as substance abuse, if any, and a detailed description of the changes in functioning that precipitated the contact. It is important to establish the nature of the onset of these changes (whether abrupt or insidious), the progression of these changes (stepwise or steady, worsening vs fluctuating vs improving), and the duration of the changes.

### **Medical conditions**

Medical history should include a review of any diseases, psychiatric or medical, and known neurological disorders, noting any history of head trauma. Alcohol or other substance abuse as well as exposure to toxins should be reviewed. Because these and other contributing factors (e.g., HIV, diabetes, urinary tract infection) may affect cognitive functioning, a careful interview documents any illness or infection, past or present.

Medication history is an important part of the initial evaluation because drug-induced cognitive changes are among the most easily reversible. All medications, including over-the-counter formulas, could have an effect on cognition, especially in combination (Albert et al., 2014; Greenblatt et al., 1991; Greenblatt, Shader, & Harmatz, 1989; Rogers, Wiese, & Rabheru, 2008). Since there has been a rise in complementary and alternative medicine, it is important to document vitamins, herbal supplements, and minerals as some could cause or interact with conditions or other medications for high blood pressure or cardiovascular conditions. Caregivers can be helpful in providing an exhaustive list of all medications being taken by the patient, complete with dosages.

A physical examination should be performed as part of the initial assessment of the geriatric patient. This part of the screening includes a brief neurological

evaluation, designed to identify lesions, vascular illness, and infection. Illnesses, such as urinary tract infection or medication toxicity, are assessed in order to rule out or address delirium. The physical examination needs to incorporate a check for signs of contusions that may indicate either accidental injuries or abuse/neglect of the patient. An evaluation should also be comprehensive for elder abuse as the prevalence of elder abuse has been reported to occur in 10% of the older population, and includes physical abuse, psychological/verbal abuse, financial exploitation, and neglect (Lachs & Pillemer, 2015). People with cognitive problems may be at increased risk as a recent study indicated that nearly 50% of individuals with dementia had sustained some form of abuse (Quinn & Benson, 2012).

### ***Family history***

A family history of dementia and other conditions (i.e., Huntington's disease and schizophrenia) should be established since the genetic component of these illnesses is significant (Bachman et al., 1993; Bierer et al., 1992; Goldberg, Ragland, & Torrey, 1990; Mayeux et al., 1993; Myers, 2004; Neale & Oltmanns, 1980; Schellenberg et al., 1992). It is important, for example, when evaluating patients who present with psychotic symptoms (i.e., delusions and hallucinations), to weigh family and personal history of schizophrenia in making a decision about the primary disorder in the clinical picture.

### ***Social adaptation***

The history-taking should include a review of educational level, career, and hobbies, along with socioeconomic, ethnic, and cultural background. All interviews should be conducted with both the patient and a caregiver, with the motives of the caregiver being assessed as well. The evaluation should take into account the possibility of either minimization or exaggeration of symptoms, depending on the family situation. Information on major life events and social supports, and especially recent changes, is necessary due to their possible contribution to the individual's performance on tests of cognitive functioning. The frequency of changes in living situations, support systems, and resources among elderly patients is common. Once an individual retires, income levels usually drop, which can require a change in living situation. The onset of medical conditions could require that one moves to a residence that provides more hands-on care, easier accessibility into the building, or smaller living quarters that are easier to maintain. In some cases, such a move may be the first one the individual has had for many years. A decrease in social supports can also occur after retirement age. One loses the stimulation of interacting with coworkers, as one may be faced with losses of both friends and family members (including spouses) that results in feelings of loneliness and sadness (Parkes, 1986; Parkes & Weiss, 1983; Zisook & Schutzer, 1986; Zisook, DeVaul, & Glick, 1982).

## ***Psychiatric conditions***

Besides evaluating family history of psychiatric conditions, the psychologist must outline the patient's own psychiatric history. If the patient does have a psychiatric history, the evaluator should ascertain whether the previous episodes were reactive or not, and what types of situations have precipitated onset of symptoms in the past.

## ***Depression***

Rates of prevalence of depression have been reported to be 11.9% for those over 71 and 13% for those 80 and over (Blazer, 2009; Steffens et al., 2009). It was also found that women and men had similar rates of depression and that Caucasians and Hispanics has almost three times the rate of depression as African Americans did (Steffens et al., 2009).

The assessment of depression in patients presenting with cognitive impairment involves some level of sophistication in order to parse out the relative contributions of affective, neurological, and other medical illness. This is critical because of treatment selection issues; if cognitive impairment is attributed to dementia, a treatable affective disorder may be overlooked. If the patient's cognitive dysfunction can be attributed with some degree of certainty to depression, the clinician has strong reasons to pursue vigorous antidepressant treatment. The failure to treat a primary depression is potentially disastrous for a patient, especially given the fairly good response of elderly depressed patients to various treatments (Alexopoulos, 2005; Benedict & Nacoste, 1990; Koenig & Blazer, 1992; Salzman & Nevis-Olesen, 1992). However, if the patient's cognitive impairment is primarily the cause of a primary dementing illness, then aggressive treatment of depressive symptoms may not substantially improve the quality of the patient's life. Overall, the issue is one of careful assessment of the clinical picture (Paquette, Ska, & Joanette, 1995).

The differential diagnosis of behavioral and cognitive disorders in older patients is made more complicated by depression, which can produce symptoms that mimic those of dementia. This is understandable given evidence from neuroimaging studies showing that patients with late-onset depression have enlarged ventricles and decreased brain density (Alexopoulos, Young, & Abrams, 1989). Estimates of the incidence of depression in the elderly indicate that it may be slightly higher among persons 65 and older than in the younger population (Blazer, 1982; Marcopoulos, 1989), and it may be the most common emotional problem among elderly patients (Hassinger, Smith, & La Rue, 1989; Thompson, Gong, Haskins, & Gallagher, 1987). Depressive symptoms are often precipitated by a traumatic loss, either of a family member, or by an event such as retirement or poor health. In cases such as these, the depression is reactive, and fits better with the diagnosis of adjustment disorder with depressed mood than with major depressive disorder. While a chronic physical illness greatly increases the likelihood of depression in an older patient, making the diagnosis of depression in a physically sick patient is often complicated by the iatrogenic factors. Depressive symptoms may arise either from an illness itself or from medication used to treat it (Greenblatt et al., 1991, 1989; Jenike, 1988).

Assessment of depression in the geriatric patient usually begins with a clinical interview of the patient, and ideally this is supplemented by corroborative information from a family member. The assessment must focus on objective symptoms of depression, including mood, behavior, anxiety, and vegetative symptoms such as sleep disturbance, anhedonia, anergia, and loss of appetite, as well as the subjective experiences outlined by the individual.

An instrument developed especially for use with elderly patients is the Geriatric Depression Scale (GDS: Brink et al., 1982; Yesavage et al., 1983). The GDS (Table 17.1) is a 30-item screening tool for depressive symptoms which has been used extensively in older adults. Factor analysis of the GDS has established a major factor of dysphoria (unhappiness, dissatisfaction with life, emptiness, downheartedness, worthlessness, helplessness) and minor factors of worry/dread/obsessive thought, and of apathy/withdrawal (Parmalee, Lawton, & Katz, 1989). Recommended cutoff points for the GDS are: normal, 0–9; mild depression,

**Table 17.1** Geriatric Depression Scale

1. Are you basically satisfied with your life?	Yes/No
2. Have you dropped many of your activities and interests?	Yes/No
3. Do you feel that your life is empty?	Yes/No
4. Do you often get bored?	Yes/No
5. Are you hopeful about the future?	Yes/No
6. Are you bothered by thoughts that you can't get out of your head?	Yes/No
7. Are you in good spirits most of the time?	Yes/No
8. Are you afraid that something bad is going to happen to you?	Yes/No
9. Do you feel happy most of the time?	Yes/No
10. Do you often feel helpless?	Yes/No
11. Do you often get restless and fidgety?	Yes/No
12. Do you prefer to stay at home rather than go out and doing new things?	Yes/No
13. Do you frequently worry about the future?	Yes/No
14. Do you feel that you have more problems with memory than most?	Yes/No
15. Do you think that it is wonderful to be alive now?	Yes/No
16. Do you often feel downhearted and blue?	Yes/No
17. Do you feel pretty worthless the way you are now?	Yes/No
18. Do you worry a lot about the past?	Yes/No
19. Do you find life very exciting?	Yes/No
20. Is it hard for you to get started on new projects?	Yes/No
21. Do you feel full of energy?	Yes/No
22. Do you feel that your situation is hopeless?	Yes/No
23. Do you think that most people are better off than you are?	Yes/No
24. Do you frequently get upset about little things?	Yes/No
25. Do you frequently feel like crying?	Yes/No
26. Do you have trouble concentrating?	Yes/No
27. Do you enjoy getting up in the morning?	Yes/No
28. Do you prefer to avoid social gatherings?	Yes/No
29. Is it easy for you to make decisions?	Yes/No
30. Is your mind as clear as it used to be?	Yes/No

*Note:* Brink et al. (1982) and Yesavage et al. (1983). Bold answers indicate depressive responses.

10–19; and severe depression, 20–30. A short form of the GDS was later developed and contains 15 items (Sheikh & Yesavage, 1986).

The common complaint of memory problems in an older adult may be associated with depression or other psychiatric disorders. Kiloh (1961) originally used the term *pseudodementia* to describe cases in which significant cognitive impairment seemed to resolve dramatically following treatment of a psychiatric condition. Because the cognitive impairment seen in depressed patients can be severe, some writers had proposed alternative terms such as *dementia syndrome of depression* (Folstein & McHugh, 1978) and *depression-related cognitive dysfunction* (Stoudemire, Hill, Gulley, & Morris, 1989). Therefore depression should be evaluated in the presence of complaints of memory problems.

Another factor in the differential diagnosis of dementia and depression is that they are often comorbid (Greenwald et al., 1989; Krishnan et al., 2002; Meyers, 1998). There is a wide range of prevalence of depression in AD with rates of 12.7% when using Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria and 42% when using other diagnostic criteria (Chi et al., 2015). Because patients' depressive symptoms may go unrecognized once AD has been diagnosed, screening should include the presence of depressive symptomatology.

## **Schizophrenia**

Older patients with cognitive impairment may exhibit psychotic symptoms. Late-life onset of psychotic symptoms may occur separately or as a secondary feature of a primary dementing condition. A differential diagnosis may become necessary to determine whether the psychotic symptoms are a component of a dementing condition like dementia of the frontal type, Huntington's disease, or AD or whether the patient is experiencing a late-onset primary psychiatric condition without dementia. While new onset psychiatric conditions in individuals with no history of psychotic disturbance is unusual, it has been reported that 23% of schizophrenic patients have an onset after 40 years old and 3% after 60 years old (Harris & Jeste, 1988). A distinction between dementias with a psychotic component and psychotic disorders without a dementing component is that late-life psychoses are typically not accompanied by profound cognitive impairments (Rosen & Zubenko, 1991; Rund, 1998). However, this area is under-researched, particularly with regard to treatment options that are safe and well-tolerated for the elderly (Nebhinani, Pareek, & Grover, 2014).

## **Cognitive functioning**

### **Definition of dementia**

Dementia is defined in several different diagnostic systems (e.g., American Academy of Neurology (AAN), 2004; American Psychiatric Association, 2013) as a condition marked by the loss of memory functions, deterioration in adaptive functioning from a higher level of functioning, and the presence of at least one additional sign of major cognitive deficit.

**Table 17.2** DSM-5 criteria for major neurocognitive disorder

Evidence of significant cognitive decline from a previous level of performance in one or more cognitive domains: <ul style="list-style-type: none"><li>• Learning and memory</li><li>• Language</li><li>• Executive function</li><li>• Complex attention</li><li>• Perceptual-motor</li><li>• Social cognition</li></ul>
The cognitive deficits interfere with independence in everyday activities. At a minimum, assistance should be required with complex instrumental activities of daily living, such as paying bills or managing medications.
The cognitive deficits do not occur exclusively in the context of a delirium.
The cognitive deficits are not better explained by another mental disorder (e.g., major depressive disorder, schizophrenia).

The American Psychiatric Association released their DSM-5 to include Major Neurocognitive disorder, replacing what was previously called dementia. The criteria are presented in [Table 17.2](#) include evidence of a cognitive decline and interferes with everyday activities. The AAN released their “Detection, Diagnosis and Management of Dementia” in 2001 and reconfirmed in 2004 outlining their guidelines on dementia. The Agency for Healthcare Research and Quality (AHRQ) has also produced guidelines for screening for dementia in primary care settings. The National Institute on Aging/Alzheimer’s Association revised their diagnostic guidelines for AD in 2011.

Though a progressive course is not necessarily a feature of dementia, many dementing conditions do entail gradual deterioration. Dementia is also distinguished from other conditions involving losses in one isolated area of cognitive function, such as amnesia. Dementia should be coded according to etiology when it can be identified. Patients who present for either medical or psychiatric evaluation may show evidence, either on examination or through complaints by either the patient himself or by a concerned relative, of the following symptoms:

*Difficulty learning new information:* Patient is repetitive, has trouble remembering recent conversations, events, and appointments; frequently misplaces objects.

*Difficulty handling complex tasks:* Patient has more trouble than expected following a complex train of thought, performing tasks that require many steps such as paying bills or preparing a meal.

*Impaired reasoning:* Patient is unable to problem-solve as effectively as in the past; shows surprising disregard for rules of social contact.

*Impaired spatial ability and disorientation:* Patient has trouble navigating in a car or with public transportation, organizing possessions in the household, or becomes confused trying to find his or her way around familiar settings.

*Language impairment:* Patient has difficulty finding words to express what he or she wants to say, and has trouble following conversations.

*Behavioral abnormalities:* Patient is less responsive and more passive, may be irritable or suspicious, may misinterpret behavior of others.

## ***Establishment of premorbid functioning***

As indicated by the discussion of normal versus impaired aging, the evaluator must establish the patient's premorbid level of cognitive functioning. Educational and occupational history will give some indication; however, we cannot dismiss the idea that some individuals can learn to cover any deficits in either of these realms. Objective measures have been shown to accurately represent premorbid functioning, including AMNART (Grober & Sliwinski, 1991; Smith, Bohac, Ivnik, & Malec, 1997), and have been utilized with older adults (Loew & Rogers, 2011). The Wechsler Test of Adult Reading (WTAR) has also been utilized to estimate premorbid intellectual functioning in older adults (Holdnack, 2001).

## ***Overview of brief dementia assessment instruments***

Along with establishing the patient's premorbid level of cognitive functioning it is often useful to administer a brief Dementia Rating Scale (DRS). The Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) is a widely used instrument intended for use as a basic preliminary screening of cognition in geriatric patients. It contains 11 cognitive tasks and can be administered in 5–10 minutes. The exam covers orientation, memory, and attention, as well as confrontation naming, praxis, and the ability to both write a sentence spontaneously and to copy overlapping pentagons. Summing the points earned for each successfully completed task produces a score of 0–30, with 30 as a perfect score. Usually the score of 23 is viewed as a threshold below which cognitive impairment is indicated (Cockrell & Folstein, 1988). However, MMSE does not measure mood, perception, nor thought content.

The Alzheimer's Disease Assessment Scale—Cognitive Subscale (ADAS; Mohs et al., 1997; Rosen, Mohs, & Davis, 1984) is widely used in clinical trials (Ihl et al., 2012). The cognitive subscale primarily measures language and memory, consists of 11 parts and takes about 30 minutes to administer. It was originally developed as a two-part scale to measure cognitive and noncognitive functions. The cognitive portion of the scale includes both short neuropsychological tests and items rated by the examiner based on both observations of the patient's behavior and an interview with the patient's caregiver. The cognitive part of the scale assesses memory, language, and praxis, while the noncognitive portion of the scale targets mood, vegetative functions, agitation, delusions, and hallucinations. The scale is designed to assess all core abnormalities, both cognitive and behavioral, that are typical of AD patients. Total scores on the cognitive subscale range from 0 to 70 and on the noncognitive subscale from 0 to 50, with increasing scores indicating greater impairment.

Clinical Dementia Rating (CDR: Hughes, Berg, Danziger, Coben, & Martin, 1982) was developed as a global measure of dementia that involves an interview with the patient and caregiver. The original CDR has been revised several times (Berg, 1988). It rates cognitive performance in six major categories: memory, orientation, judgment and problem-solving, community affairs, home and hobbies, and personal care. Scores on these six ratings are synthesized into a single score

categorizing the stage of dementia ranging from none (0), to questionable (0.5), mild (1), moderate (2), or severe (3).

The DRS-2 ([Jurica, Leitten, & Mattis, 2001](#)) is somewhat more comprehensive than some of the briefer scales such as the CDR and the MMSE. The DRS-2 consists of five subscales: attention, initiation-perseveration, construction, conceptualization, and memory. A score on the DRS ranges from 0 to 144. Normal scores are 140 and above, and a DRS score under 100 indicates severe impairment.

The usefulness of brief DRS is clear when a gross measure of functioning is needed, either for screening or research. Brief DRS that have multiple forms, such as the ADAS, are particularly useful in outcome studies, when the information sought is objective change. However, more extensive neuropsychological batteries are often indicated in both research and clinical practice. This is because of the additional information that can be gleaned regarding an individual's specific cognitive strengths and weaknesses, in order to identify predictors of the course of a particular illness and to help differentiate between different subsets of certain psychiatric disorders. In addition, neuropsychological data is useful in the development of treatment strategies tailored to the pattern of individual strengths and weaknesses demonstrated on testing ([Keefe, 1995](#)).

## **General principles in neuropsychological assessment of the older adult**

Aging itself is associated with changes in virtually every function that becomes impaired in dementia. Most of the cognitive deficits seen in dementing illnesses are essentially exaggerations of normal age-related changes. Only when these deficits exceed expected levels for the patient's age and educational group, when those deficits affect adaptation, or when psychiatric symptoms (delusions, hallucinations, depression) occur, do cognitive deficits fall into the realm of a dementia. Those functions that are most resistant to decline in normal aging, such as word knowledge, are also preserved longest in most dementias.

The initial assessment of a geriatric patient is complicated by several factors including the patient's age, premorbid intelligence and previous level of functioning, educational attainment, cultural background, comorbid psychiatric illness, sensory deficits, and medical status. Thus these factors must be considered when working with patients. Once symptoms of a possible dementia have been recognized, a thorough assessment should be initiated. This assessment consists of a detailed history, physical examination, and neuropsychological assessment of functional ability and mental status. Neuroimaging is indicated very often.

The neuropsychological assessment is an actuarial approach to the quantification of impairments reported by the patient, such as those mentioned above. The subjective complaints of the patient translate into cognitive domains targeting for evaluation, including perception, attention, learning and memory, verbal skills, motor skills, and executive function.

A neuropsychological test battery for the geriatric patient should begin with a thorough history, as discussed above. The primary purpose of the history is to establish a strong foundation on which to base estimates of a patient's premorbid level of functioning. There would be very different assumptions drawn about the premorbid level of functioning of individuals testing in the average range on the WAIS-IV, depending on the history. For example, if that person was known to have only a 10th-grade formal education, and worked as a custodian off and on throughout his life, it would be reasonable to generate the hypothesis that that individual may have had some psychiatric problems that interfered in his ability to function at a level comporting with his intellectual capacity. On the other hand, if the individual being tested had achieved a doctoral degree and had functioned until her recent retirement as chairperson of an academic department at a university, an average performance on the WAIS-IV would suggest a recent intellectual deterioration.

In order to establish a baseline, or premorbid level of functioning, it is useful to estimate from performance on tests of old learning, because of the minimal effect of aging on such tests. All cognitive impairments identified in geriatric patients must be referenced to age-corrected norms as well as to the patient's previous level of cognitive functioning, as noted above.

### ***Learning and memory***

Memory problems are a frequent complaint in the aging population. The memory domain includes the ability to retain information for a very brief period (primary or working memory), encode information for transfer to long-term storage, acquire information with repeated trial exposure (serial learning), retrieve information from memory after a delay, either with or without cues (delayed recognition vs recall), resistance to interference during the retention interval, and the ability to retrieve information that was learned long ago and bring it into current usage (long-term memory). Many of these processes are examined in the typical neuropsychological examination, with slightly different patterns of impairment depending on the etiology of the cognitive impairments in question.

### ***Attention***

Attention is a construct about which there is considerable controversy. In general, this construct refers to the ability of individuals to identify (register), focus on, process, and sustain contact with information to the extent that other operations can be performed on it. There is, however, substantial overlap between attentional processes and others that are labeled perception and memory. For example, an object must be recognized at the same time it is being perceived. Working memory, the ability to sustain information in memory while it is being processed, interacts with sustained attention. Sustained attention necessitates, in turn, that the object is maintained in working memory. Regardless of these

interactions, attentional skills usually deteriorate in a broad sense in various dementing conditions. Many studies that have focused on attention target both verbal and spatial stimuli.

### ***Perception***

Perception is the ability to identify objects and other information on the basis of interpretation of sensory input. Each of the five senses is involved in this process and each may potentially be impaired in certain dementing conditions. Structured tests are used to examine each, although the majority of the attention has been focused on visual and tactile functions.

### ***Verbal skills***

This area of functioning refers to the ability to use language adaptively, both expressively and receptively. Generally, demented patients have difficulty with generating coherent speech, with reduced complexity and content. In the assessment of verbal skills in dementia, several aspects of functioning have received considerable attention. Fluency, the ability to consistently produce words in response to a situational or task demand, has been closely examined, as has the ability to verbally identify objects (confrontation naming). In addition, vocabulary, reading ability, and other well-learned verbal skills, including the ability to use appropriate grammar and syntax, are also often affected by certain types of dementing disorders. Since deficits in receptive language ability can result in a difficulty expressing oneself, identification of specific verbal impairments is important to accomplish during the course of a neuropsychological evaluation.

### ***Motor skills***

Motor skills can be simple, such as opening a door using a doorknob, or much more complex, for example reproducing a complex drawing or performing a sonata on the piano. Some motor skills tasks require an external stimulus, such as a model that is copied by the subject. These tasks are viewed as tapping “constructional praxis.”

### ***Executive functioning***

This domain refers to the ability to plan and organize behavior, to process more than one simultaneous stream of information, and to coordinate the application of several cognitive resources to accommodate competing demands. Executive functioning is comprised of abilities such as planning, reasoning, initiation, inhabitation, set shifting, abstraction, and other higher order abilities. Deficits in executive functioning tasks can be the result of deficits in any one of the other cognitive domains, or in their integration.

## **Praxis**

Praxis refers to deliberate control of the motor skills employed in the execution of complex learned movements. It is usually tested by giving the patient a series of commands to follow, from simple (*pretend to comb your hair*) through facial (*whistle*) to more complex (*address a letter to yourself; pretend to knock at the door and open it*). Praxis is often subsumed under other categories such as constructional abilities, which include the ability to construct figures according to verbal directions (e.g., draw a clock) and is related to visuomotor integrative skills, such as the ability to copy a figure in two-dimensional space (e.g., Bender–Gestalt test; [Bender, 1938](#)) or three-dimensional space (e.g., Block Design subtest of the WAIS-IV).

## **Visuospatial organization**

Related to the above are visuointegrative skills, defined as the ability to put together pieces of a puzzle so that they form a whole (Block Design, Hooper Visual Organization Test). Disorders of praxis and visuospatial organization tend to go together, though they can be seen in isolation. Thus an individual may experience difficulty with all constructional tasks, or may be able to copy well but not be able to perform mental rotations of parts to create a whole percept. Others may be able to perform mental rotations but not be able to organize a complex drawing on paper (Rey Osterrieth Complex Figure; [Osterrieth, 1944](#); [Rey, 1941](#)).

There are many texts available that describe the abovementioned cognitive domains in greater detail (e.g., [Lezak et al., 2012](#)), and which provide a comprehensive list of neuropsychological tests (Strauss et al., 2006).

## **Differential diagnosis**

### **Profile analysis**

In analyzing data from neuropsychological evaluations, it is clear that different profiles emerge for different patients, depending on the etiology of the complaint. We will limit our discussion of typical profiles here to those most commonly encountered in geriatric neuropsychology; namely, dementia and depression. The typical neuropsychological referral in geropsychological practice is generated when patients complain to their psychiatrist or internist about cognitive deterioration. While the competent clinician can confirm by using a brief screening measure such as the Mini-Mental State Exam (MMSE; [Folstein et al., 1975](#)) that cognitive changes have occurred, more in-depth evaluation is indicated. Neuropsychological testing reveals differing patterns of scores that may be helpful in distinguishing the etiology of the cognitive disturbance, shedding light on strengths and weaknesses that will potentially affect the development of treatment plans for the patient.

## **Alzheimer's disease**

This disease, first reported in the early part of this century, is the most common cause of dementia. It is estimated that 5.4 million Americans have AD and by 2050 that number will grow to 13.8 million (Alzheimer's Association, 2016). In 2013 AD was the fifth leading cause of death in individuals aged 65 and older and while the rates of death from stroke, heart disease and prostate cancer decreased, deaths from AD increased 71% between 2000 and 2013 (Alzheimer's Association, 2016). The course of AD is about 10 years from the first sign of illness. Risk factors include age, family history of AD, low education, and rural residence (Hall et al., 1998). On autopsy, the brain is found to have amyloid plaques and neurofibrillary tangles. These abnormalities are initially located in the medial temporal cortex and hippocampus, and eventually spread to the temporal lobe, parietal cortex, and the frontal lobe.

AD is distinguished from other dementias by a deteriorating course. The first indication of an AD dementia is a profound deficit in serial learning and delayed recall. This corresponds with the very common presenting complaint of the patient: forgetfulness and difficulty learning new material. This deficit is profound, and is apparent on neuropsychological testing even in patients who may have virtually normal MMSE scores. These patients, though scoring in the mildly impaired or better range on the MMSE, will perform much more poorly than expected on tests of delayed recall (such as the Logical Memory and Visual Reproduction subtests of the WMS-IV) and serial learning (e.g., Rey Auditory-Verbal Learning Test; California Verbal Learning Test—2). On tests of delayed recall, AD patients display virtually no retention, compared with over 85% retained by normal adults (Welsh, Butters, Hughes, Mohs, & Heyman, 1991). Opportunities to rehearse new material do not seem to benefit AD patients, nor does cueing, though these conditions allow normals to improve their performance on memory tests (Weingartner et al., 1993).

Impairments in memory and learning are followed by deficits in verbal skills. In particular, category fluency as measured by a test such as Animal Naming, has been shown to be an early hallmark of Alzheimer's type dementia (Bayles et al., 1989; Butters, Granholm, Salmon, Grant, & Wolfe, 1987; Monsch et al., 1992; Pasquier, Lebert, Grymonpre, & Perit, 1995) while phonemic fluency does not decline until later in the course. This has been shown to be related to a deficit in semantic knowledge that affects relationships among lower level concepts, more so than the relationship between the concepts and their higher order category of membership (Glosser, Friedman, Grugan, Lee, & Grossman, 1998).

Executive functioning deficits appear early in AD, with confrontation naming, praxis, and visuospatial deficits appearing later and progressing in a linear fashion. Motor speed is impaired early in the illness, and gets progressively worse (Nebes & Brady, 1992). Attention and concentration is intact early, though orientation is often impaired initially. As the disease progresses, concentration declines gradually (Kaszniak, Poon, & Riege, 1986). AD progresses steadily until performance on all tests reaches the floor (Zec, 1993). Many patients suffer from behavioral and mood disturbances, including delusions, hallucinations, agitation, and depression.

## **Vascular dementia**

Vascular dementia is the second most common dementia after AD (Jorm, 1991) and impacts 1%–4% of individuals over 65 (Hebert & Brayne, 1995). Because stroke can affect any and all regions of the brain, there is no single profile for cognitive impairment caused by vascular disease. Patients who have vascular dementia as well as those with mixed dementia (AD and vascular) have been found to have deficits in memory, orientation, language, and concentration and attention with the only marked difference between the two groups the presence of gait disturbance and lesser impairments in naming and praxis among those with vascular dementia alone (McVeigh & Passmore, 2006; Thal, Grundman, & Klauber, 1988).

Vascular dementia patients are also seen as displaying a pattern of “patchy” or irregular deficits, with clear deficits that do not follow any pattern across patient groups. A demented patient whose deficits are predominantly in the area of executive functioning would be likely to have suffered infarction in the frontal lobes, while a demented patient with aphasia may have strokes in the frontotemporal region. Subcortical vascular dementias are often characterized by profound slowing of movement (bradykinesia) and thought (bradyphrenia) such as that in the subcortical dementias associated with Parkinson’s and Huntington’s diseases.

In the assessment of vascular dementias it is also important to get a thorough history of the course of the impairment. A single stroke may lead to a focal pattern of impairment, in which memory is largely unscathed, to a diffuse pattern in which memory and other domains are affected. While AD is characterized by a persistent deteriorating course, vascular dementia is traditionally “stepwise” in its pattern of decline (Hachinski, Lassen, & Marshall, 1974). There has been some indication of recovery of cognitive function in patients following treatment of vascular disease (Hershey, Modic, Jaffe, & Greenough, 1986) just as there is often continued mental decline related to additional infarction.

There have been contributions to the field that indicate that infarction is not the only vascular condition that may lead to cognitive changes. White matter disease has also been associated with a dementia syndrome that has particular impact on frontal lobe abilities such as executive function, attention, and overall intellectual level, with relative sparing of language, memory, and visual–spatial skills (Boone, Miller, & Lesser, 1993; Libon et al., 1997). Other studies point to additional types of vascular disease, such as atrophy, gliosis and spongiosis (Gustafson, 1987), white and gray matter changes (Gydesen, Hagen, & Klinken, 1987; Libon et al., 1997), atrophy and gliosis (Neary, Snowden, & Mann, 1990), all of which may lead to cognitive impairment.

## **Depression versus dementia**

The differential diagnosis between depressed and dementia is complex but may be made using a combination of a neuropsychological evaluation and a thorough mood assessment. Although symptoms of depression may be due a mood disorder, they may also be due to early signs of dementia (Bieliauskas, 2012; Saczynski et al., 2010).

There is not a consensus in the literature regarding the mechanisms to determine if depression is a prodrome or risk for dementia (Wright & Persad, 2007). **Table 17.3** summarizes the differences in neuropsychological performance between individuals with depression and dementia.

**Table 17.3** Depression versus dementia

	<b>Depression</b>	<b>Early Alzheimer's disease</b>
<i>Cognitive function</i>		
Memory		
Recognition	Relatively intact	Impaired
Immediate	Mild attentional difficulties	Moderately to severely impaired
Delayed recall	Near normal rate	Little to no retention
Learning		
New information	Intact	Severely impaired
Complex tasks	Distractible	Loses train of thought easily
Reasoning	Intact	Impaired
Attention		
Perception		
Language skills		
	Normal expressive and receptive functioning; reduced verbal fluency	Decline in expressive and receptive functioning
Executive functioning	Intact	Mild impairments evident early, especially parallel processing
Praxis	Slowed	Intact
Visuospatial	Normal	Impaired
<i>Course of illness</i>		
Onset	Rapid	Insidious
Awareness of impairment	Intact, complaints of memory problems	Impaired
Duration	Few weeks to months; reversible with treatment	Deteriorating course over approximately 10 years
<i>Mood</i>		
Symptoms	Stable level of depression, apathy, and withdrawal	Labile—between normal and withdrawn
<i>Somatic</i>		
Symptoms	Vegetative signs: insomnia, eating disturbances, minor physical complaints	Some sleep disturbances

## Special problems in geriatric assessments

While it is usually interesting to do a follow-up assessment on psychiatric patients to evaluate changes in functioning since a previous evaluation, reexamination is a virtual necessity in geropsychology. There are several times when retesting is especially important. Patients with little formal education and/or a borderline or lower IQ may initially perform at such low levels on neuropsychological testing that it is extremely difficult to distinguish a dementia from baseline performance. In such cases, retesting is necessary in order to establish presence or absence of a deteriorating course. Even though patients may be performing at an extremely low percentile level when first tested, it is possible to discern changes over time in the raw scores. If the raw scores drop noticeably and consistently across the different domains tested, even while changes in percentile level are not discernable because of an extremely low baseline, it is possible to conclude that there has been a global deterioration over time. A deteriorating course is a hallmark of AD; in order to establish the existence of such a course in patients with extremely low baseline performance, reexamination is necessary.

At the other end of the intelligence spectrum, and presenting another diagnostic conundrum, are those elderly patients who have extremely high levels of intellectual functioning. Notwithstanding their advanced age, individuals with IQs in the upper reaches of the scale (Superior and Very Superior ranges) often present themselves for evaluation because of subjective complaints of memory loss. These individuals, accustomed to enjoying great mental acuity, may be particularly sensitive to any diminishment of their abilities. They will often hit the ceiling on a gross screening measure such as the Mini-Mental Status examination, achieving perfect or near-perfect scores. Neuropsychological assessment will be necessary to determine whether these patients have suffered significant cognitive losses, or whether they are performing at expected levels. While the profiles of two such patients may be similar in that their premorbid levels of functioning are in the Very Superior range, the delayed recall performance will differentiate a patient with early dementia from another who has suffered some changes in functioning but whose memory performance is still within the Very Superior range for her age. The patient who is not demented will have a relatively lower raw score than he may have achieved on previous testing, but his age-scaled score remains essentially the same. Another patient with a premorbid IQ of 150 with early AD may have delayed recall scaled scores as high as 50th percentile: this is still in the normal range but represents a significant deficit for him.

Regardless of whether the patient in question is at one or the other end of the IQ curve, the extremity of their scores dictates that retesting will play a critical role in the assessment process. The first testing is necessary to establish a baseline, and the second, usually 1 year later, will be useful in determining a course.

## References

- Adams, K. M. (1986). Concepts and methods in the design of automata for neuropsychological test interpretation. In S. B. Filshov, & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (Vol. 2). New York: John Wiley & Sons.
- Adams, K. M., & Heaton, R. K. (1985). Automated interpretation of neuropsychological test data. *Journal of Consulting and Clinical Psychology*, 53, 790–802.
- Albert, S. M., Bix, L., Bridgeman, M. M., Carstensen, L. L., Dyer-Chamberlain, M., Neafsey, P. J., & Wolf, M. S. (2014). Promoting safe and effective use of OTC medications: CHPA-GSA national summit. *The Gerontologist*, 54(6), 909–918.
- Alexopoulos, G. S. (2005). Depression in the elderly. *The Lancet*, 365, 1961–1970.
- Alexopoulos, G. S., Young, R. C., Abrams, R. C., et al. (1989). Chronicity and relapse in geriatric depression. *Biological Psychiatry*, 26, 551–564.
- Alzheimer's Association. (2016). 2016 Alzheimer's disease facts and figures. *Alzheimer's Dementias*, 12(4), 459–509.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders, fifth edition (DSM-5)*. Arlington, VA: American Psychiatric Association.
- Bachman, D. L., Wolf, P. A., Linn, R. T., Knoefel, J. E., Cobb, J. L., Belanger, A. J., ... D'Agostino, R. B. (1993). Incidence of dementia and probable Alzheimer's disease in a general population: The Framingham Study. *Neurology*, 43, 515–519.
- Bayles, K. A., Salmon, D. P., Tomoeda, C. K., Jacobs, D., Caffrey, J. T., Kaszniak, A. W., & Troster, A. I. (1989). Semantic and letter category naming in Alzheimer's patients: A predictable difference. *Developmental Neuropsychology*, 5, 335–347.
- Bender, L. (1938). *Instructions for the use of the Bender visual-motor Gestalt test*. American Orthopsychiatric Association.
- Benedict, K. B., & Nacoste, D. B. (1990). Dementia and depression: A framework for addressing difficulties in differential diagnosis. *Clinical Psychology Review*, 10, 513–537.
- Berg, L. (1988). Mild senile dementia of the Alzheimer's type: Diagnostic criteria and natural history. *Mount Sinai Journal of Medicine*, 55, 87–96.
- Bierer, L. M., Silverman, J. M., Mohs, R. C., Haroutunian, V., Li, G., Purohit, D., ... Davis, K. L. (1992). Morbid risk to first degree relatives of neuropathologically confirmed cases of Alzheimer's disease. *Dementia*, 3, 134–139.
- Blazer, D. (1982). The epidemiology of late life depression. *Journal of the American Geriatrics Society*, 30, 587–592.
- Blazer, D. G. (2009). Depression in late life: Review and commentary. *FOCUS*, 7(1), 118–136.
- Boone, K. B., Miller, B. L., & Lesser, I. M. (1993). Frontal lobe cognitive functions in aging: Methodological considerations. *Dementia*, 4, 232–236.
- Bowler, C., Boyle, A., Branford, M., Cooper, S. A., Harper, R., & Lindesay, J. (1994). Detection of psychiatric disorders in elderly medical inpatients. *Age and Ageing*, 23(4), 307–311.
- Brink, T. L., Yesavage, J. A., Lum, O., Heersema, P. H., Adey, M., & Rose, T. S. (1982). Screening tests for geriatric depression. *Clinical Gerontologist*, 1, 31–43.
- Bucks, R. S., Ashworth, D. L., Wilcock, G. K., & Siegfried, K. (1996). Assessment of activities of daily living in dementia: development of the Bristol Activities of Daily Living Scale. *Age and Ageing*, 25, 113–120. Available from <https://doi.org/10.1093/ageing/25.2.113>.

- Butters, N., Granholm, E., Salmon, D. P., Grant, I., & Wolfe, J. (1987). Episodic and semantic memory: A comparison of amnesic and demented patients. *Journal of Clinical Neuropsychology, 9*, 479–497.
- Cassel, J. (1976). The contribution of social environment to host resistance: The fourth Wade Hampton Frost lecture. *American Journal of Epidemiology, 104*, 107–123.
- Chi, S., Wang, C., Jiang, T., Zhu, X. C., Yu, J. T., & Tan, L. (2015). The prevalence of depression in Alzheimer's disease: A systematic review and meta-analysis. *Current Alzheimer Research, 12*, 189–198.
- Christensen, A. L. (1979). *Luria's neuropsychological investigation* (2nd ed.). Copenhagen: Munksgaard.
- Cockrell, J. R., & Folstein, M. F. (1988). Mini-Mental State Examination (MMSE). *Psychopharmacology Bulletin, 24*, 689–692.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state. *Journal of Psychiatric Research, 12*, 189–198.
- Folstein, M. F., & McHugh, P. R. (1978). Dementia syndrome of depression. *Aging, 7*, 87–93.
- Glass, T. A., Seeman, T. E., Hertzog, A. R., Kahn, R. L., & Berkman, L. F. (1995). Changes in productivity in late adulthood: MacArthur studies of successful aging. *Journal of Gerontology: Social Sciences, 50B*, S65–S76.
- Glosser, G., Friedman, R. B., Grugan, P. K., Lee, J. H., & Grossman, M. (1998). Lexical semantic and associative priming in Alzheimer's disease. *Neuropsychology, 12*(2), 218–224.
- Goldberg, T. E., Ragland, J. D., Torrey, E. F., et al. (1990). Neuropsychological assessment of monozygotic twins discordant for schizophrenia. *Archives of General Psychiatry, 47*, 1066.
- Greenblatt, D. J., Harmatz, J. S., Shapiro, L., Engelhardt, N., Gouthro, T. A., & Shader, R. I. (1991). Sensitivity to triazolam in the elderly. *New England Journal of Medicine, 324*, 1691–1698.
- Greenblatt, D. J., Shader, R. I., & Harmatz, J. S. (1989). Implications of altered drug disposition in the elderly: Studies of benzodiazepines. *Journal of Clinical Pharmacology, 29*, 866–872.
- Greenwald, B. S., Kramer-Ginzberg, E., Marin, D. B., Laitman, L. B., Herman, C. K., Mohs, R. C., & Davis, K. L. (1989). Dementia with coexisting depression. *American Journal of Psychiatry, 146*, 1472–1478.
- Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology, 13*, 933–949.
- Gustafson, L. (1987). Frontal lobe degeneration of the non-Alzheimer type, II: Clinical picture and differential diagnosis. *Archives of Gerontology and Geriatrics, 6*, 209–224.
- Gydesen, S., Hagen, S., Klinken, L., et al. (1987). Neuropsychiatric studies in a family with presenile dementia different from Alzheimer's and Pick's disease. *Acta Psychiatry Scandinavica, 76*, 276–284.
- Hachinski, V. C., Lassen, N. A., & Marshall, J. (1974). Multi-infarct dementia: A cause of mental deterioration in the elderly. *Lancet, 2*, 207–209.
- Hall, K., Unverzagt, F. W., Hendrie, H. C., Gurje, O., Gao, S., Hui, S. L., & Baiyewu, O. (1998). Risk factors and Alzheimer's disease: A comparative study of two communities. *Australian & New Zealand Journal of Psychiatry, 32*, 698–706.
- Harris, M. J., & Jeste, D. V. (1988). Late-onset schizophrenia: An overview. *Schizophrenia Bulletin, 14*, 39–55.

- Harvey, P. D., & Dahlman, K. L. (1998). Neuropsychological evaluation of dementia. In A. Chalev (Ed.), *Neuropsychological assessment of neuropsychiatric disorders*. Washington, DC: American Psychiatric Press.
- Harwood, D. M. L., Hope, T., & Jacoby, R. (1997a). Cognitive impairment in medical inpatients. I: Screening for dementia—Is history better than mental state? *Age and Ageing*, 26, 31–35.
- Harwood, D. M. L., Hope, T., & Jacoby, R. (1997b). Cognitive impairment in medical inpatients. II: Do physicians miss cognitive impairment? *Age and Ageing*, 26, 37–39.
- Hassinger, M., Smith, G., & La Rue, A. (1989). Assessing depression in older adults. In T. Hunt, & C. J. Lindley (Eds.), *Testing older adults: A reference guide for geropsychological assessments*. Austin, TX: Pro-Ed.
- Hebert, R., & Brayne, C. (1995). Epidemiology of vascular dementia. *Neuroepidemiology*, 14, 240–257.
- Hershey, L. A., Modic, M. T., Jaffe, D. F., & Greenough, P. G. (1986). Natural history of the vascular dementia: A prospective study of seven cases. *Canadian Journal of Neurological Sciences*, 13, 559–565.
- Holdnack, H. A. (2001). *Wechsler test of adult reading: WTAR*. San Antonio, TX: The Psychological Corporation.
- House, J. S., Landis, K. R., & Umberson, D. (1988). Social relationships and health. *Science*, 241, 540–545.
- Hughes, C., Berg, L., Danziger, W. L., Coben, L. A., & Martin, R. L. (1982). A new clinical scale for staging of dementia. *British Journal of Psychiatry*, 140, 566–572.
- Ihl, R., Ferris, S., Robert, P., Winblad, B., Gauthier, S., & Tennigkeit, F. (2012). Detecting treatment effects with combinations of the ADAS-cog items in patients with mild and moderate Alzheimer's disease. *International Journal of Geriatric Psychiatry*, 27(1), 15–21.
- Jenike, M. A. (1988). Depression and other psychiatric disorders. In M. S. Albert, & M. Moss (Eds.), *Geriatric neuropsychology* (pp. 115–144). New York: Guilford Press.
- Jorm, A. F. (1991). Cross-national comparisons of the occurrence of Alzheimer's and vascular dementias. *European Archives of Psychiatry and Clinical Neuroscience*, 240, 218–222.
- Jurica, P. J., Leitten, C. L., & Mattis, S. (2001). *Dementia Rating Scale-2*. Odessa, FL: Psychological Assessment Resources.
- Kahn, R. L., & Byosiere, P. (1992). Stress in organizations. In (2nd ed.) M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3) Palo Alto, CA: Consulting Psychologists Press.
- Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll, & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function: Research, measurement, and practice*. Washington, DC: American Psychological Association.
- Kasznaik, A. W., Poon, L. W., & Riege, W. L. (1986). Assessing memory deficits: An information-processing approach. In L. W. Poon (Ed.), *Clinical memory assessment of older adults*. Washington, DC: American Psychological Association.
- Katzman, R., Lasker, B., & Berstein, N. (1988). Advances in the diagnosis of dementia: Accuracy of diagnosis and consequences of misdiagnosis of disorders causing dementia. In R. D. Terry (Ed.), *Aging and the brain* (pp. 17–62). New York: Raven Press.
- Keefe, R. S. E. (1995). The contribution of neuropsychology to psychiatry. *American Journal of Psychiatry*, 152, 6–15.
- Kiloh, L. G. (1961). Pseudo-dementia. *Acta Psychiatrica Scandinavica*, 37, 336–351.
- Koenig, H. G., & Blazer, D. G. (1992). Mood disorders and suicide. In J. E. Birren, R. B. Sloane, & G. D. Cohen (Eds.), *Handbook of mental health and aging* (2nd ed., pp. 379–407). San Diego, CA: Academic Press.

- Krishnan, K. R., Delong, M., Kraemer, H., Carney, R., Spiegel, D., Gordon, C., ... Wainscott, C. (2002). Comorbidity of depression with other medical diseases in the elderly. *Biological Psychiatry*, 52(6), 559–588.
- Lamberty, G. J., & Bieliuskas, L. A. (1993). Distinguishing between depression and dementia in the elderly: A review of neuropsychological findings. *Archives of Clinical Neuropsychology*, 8, 149–170.
- Lezak, M. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York: Oxford University Press.
- Libon, D. J., Bogdanoff, B., Bibavuta, J., Skalina, S., Cloud, B. S., Resh, R., ... Ball, S. K. (1997). Dementia associated with periventricular and deep white matter alterations: A subtype of subcortical dementia. *Archives of Clinical Neuropsychology*, 12(3), 239–250.
- Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books.
- Luria, A.R. (1973). *The working brain: An introduction to neuropsychology* [Trans. B. Haigh]. New York: Basic Books.
- Marcopulos, B. A. (1989). Pseudodementia, dementia, and depression: Test differentiation. In T. Hunt, & C. J. Lindley (Eds.), *Testing older adults: A reference guide for geropsychological assessments*. Austin, TX: Pro-Ed.
- Mayeux, R., Ottman, R., Tang, M. X., NoboaBauza, L., Marder, K., Gurland, B., & Stern, Y. (1993). Genetic susceptibility and head injury as risk factors for Alzheimer's disease among community dwelling elderly persons and their first degree relatives. *Annals Neurology*, 33, 494–501.
- McVeigh, C., & Passmore, P. (2006). Vascular dementia: prevention and treatment. *Clinical Interventions in Aging*, 1(3), 229–235.
- Meyers, B. S. (1998). Depression and dementia: Comorbidities, identification, and treatment. *Journal of Geriatric Psychiatry Neurology*, 11(4), 201–205.
- Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., et al. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. The Alzheimer's Disease Cooperative Study. *Alzheimer Disease & Associative Disorders*, 11(Suppl. 2), S13–S21.
- Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., & Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of Neurology*, 49, 1253–1258.
- Myers, R. H. (2004). Huntington's disease genetics. *NeuroRx*, 1(2), 255–262.
- Neale, J. M., & Oltmanns, T. F. (1980). *Schizophrenia*. New York: John Wiley.
- Neary, D., Snowden, J. S., Mann, D. M. A., et al. (1990). Frontal lobe dementia and motor neuron disease. *Journal of Neurology Neurosurgery Psychiatry*, 53, 23–32.
- Nebes, R. D., & Brady, C. B. (1992). Different patterns of cognitive slowing produced by Alzheimer's disease and normal aging. *Journal of Clinical and Experimental Neuropsychology*, 14, 317–326.
- Nebhinani, N., Pareek, V., & Grover, S. (2014). Late-life psychosis: An overview. *Journal of Geriatric Mental Health*, 1, 60–70.
- Osterrieth, P. A. (1944). Le test de copie d'une figure complexe: Contribution à l'étude de la perception et de la memoire. *Archives de Psychologie*, 30, 286–356.
- Paquette, I., Ska, B., & Joanette, Y. (1995). Delusions, hallucinations, and depression in a population-based, epidemiological sample of demented subjects. In M. Bergener, & S. I. Finkel (Eds.), *Treating Alzheimer's and other dementias*. New York: Springer Publishing Co.

- Parkes, C. M. (1986). *Bereavement: Studies of grief in adult life*. Madison CT: International Universities Press.
- Parkes, C. M., & Weiss, R. S. (1983). *Recovery from bereavement*. New York: Basic Books.
- Parmalee, P. A., Lawton, M. P., & Katz, I. R. (1989). Psychometric properties of the Geriatric Depression Scale among the institutionalized aged. *Psychological Assessment*, 1, 331–338.
- Pasquier, F., Lebert, F., Grymonpre, L., & Perit, H. (1995). Verbal fluency in dementia of the frontal lobe type and dementia of Alzheimer's disease. *Journal of Neurology Neurosurgery Psychiatry*, 58, 81–84.
- Reitan, R. M., & Davidson, L. A. (1974). *Clinical neuropsychology: Current status and applications*. New York: Winston/Wiley.
- Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.
- Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives de Psychologie*, 28, 286–340.
- Rogers, J., Wiese, B. S., & Rabheru, K. (2008). The older brain on drugs: Substances that may cause cognitive impairment. *Geriatrics and Aging*, 11(5), 284–289.
- Rosen, J., & Zubenko, G. S. (1991). Emergence of psychosis and depression in the longitudinal evaluation of Alzheimer's disease. *Archives of Geriatric Gerontology*, 6, 225–233.
- Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *American Journal of Psychiatry*, 141, 1356–1364.
- Rowe, J. W., & Kahn, R. L. (1987). Usual and successful aging. *Science*, 237, 143–148.
- Rowe, J. W., & Kahn, R. L. (1997). Successful aging. *Gerontologist*, 37, 433–440.
- Russell, E. W., Neuringer, C., & Goldstein, G. (1970). *Assessment of brain damage: A neuropsychological key approach*. New York: Wiley-Interscience.
- Saczynski, J. S., Beiser, A., Seshadri, S., Auerbach, S., Wolf, P. A., & Au, R. (2010). Depressive symptoms and risk of dementia. *Neurology*, 75(10), 35–41.
- Salzman, C., & Nevis-Olesen, J. (1992). Psychopharmacologic treatment. In J. E. Birren, R. B. Sloane, & G. D. Cohen (Eds.), *Handbook of mental health and aging* (2nd ed., pp. 722–762). San Diego, CA: Academic Press.
- Samuels, S. C., & Davis, K. L. (1998). Use of cognitive enhancers in dementing disorders. In J. C. Nelson (Ed.), *Geriatric psychopharmacology* (pp. 381–403). New York: Marcel Dekker, Inc.
- Schellenberg, G. D., Bird, T. D., Wijsman, E. M., Orr, H. T., Anderson, L., Nemens, E., ... Martin, G. M. (1992). Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science*, 258, 668–671.
- Sheikh, J. I., & Yesavage, J. A. (1986). Geriatric Depression Scale (GDS). Recent evidence and development of a shorter version. In T. L. Brink (Ed.), *Clinical gerontology: A guide to assessment and intervention* (pp. 165–173). NY: The Haworth Press, Inc.
- Smith, G. E., Bohac, D. L., Ivnik, R. J., & Malec, J. F. (1997). Using word recognition tests to estimate premorbid IQ in early dementia: Longitudinal data. *Journal of the International Neuropsychological Society*, 3, 528–533.
- Stoudemire, A., Hill, C., Gulley, L. R., & Morris, R. (1989). Neuropsychological and biomedical assessment of depression–dementia syndromes. *Journal of Neuropsychiatry and Clinical Neurosciences*, 1, 347–361.
- Swiercinsky, D. P. (1978). *Manual for the adult neuropsychological evaluation*. Springfield, IL: C.C. Thomas.
- Thal, L. J., Grundman, M., & Klauber, M. R. (1988). Dementia: Characteristics of a referral population and factors associated with progression. *Neurology*, 38, 1083–1090.

- Thompson, L. W., Gong, V., Haskins, E., & Gallagher, D. (1987). Assessment of depression and dementia during the late years. In K. W. Schaie (Ed.), *Annual review of gerontology and geriatrics*. New York: Springer.
- Wechsler, D. (1997a). *Wechsler adult intelligence scale manual* (3rd ed.). San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997b). *Wechsler memory scale manual* (3rd ed.). San Antonio, TX: The Psychological Corporation.
- Weingartner, H. R., Eckart, M., Grafman, J., Molchan, S., Putnam, K., Rawlings, R., & Sunderland, T. (1993). The effects of repetition on memory performance in cognitively impaired patients. *Neuropsychology*, 7, 385–395.
- Welsh, K. A., Butters, N., Hughes, J., Mohs, R. C., & Heyman, A. (1991). Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. *Archives of Neurology*, 48, 278–281.
- Williams, J. M. (1997). The prediction of premorbid memory ability. *Archives of Clinical Neuropsychology*, 12, 745–756.
- Wright, S. L., & Persad, C. (2007). Distinguishing between depression and dementia in older persons: Neuropsychological and neuropathological correlates. *Journal of Geriatric Psychiatry and Neurology*, 20(4), 189–198.
- Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M. B., & Leirer, V. O. (1983). Development and validation of a geriatric depression rating scale: A preliminary report. *Journal of Psychiatric Research*, 17, 37–49.
- Zec, R. F. (1993). Neuropsychological functioning in Alzheimer's disease. In R. W. Parks, R. F. Zec, & R. S. Wilson (Eds.), *Neuropsychology of Alzheimer's disease and related disorders*. New York: Oxford University Press.
- Zisook, S., DeVaul, R. A., & Glick, M. A. (1982). Measuring symptoms of grief and bereavement. *American Journal of Psychiatry*, 139, 1590–1593.
- Zisook, S., & Schueter, S. R. (1986). The first four years of widowhood. *Psychiatric Annals*, 15, 288–294.

## Further reading

- Albert, M. S., & Moss, M. B. (1988). *Geriatric neuropsychology*. New York: Guilford.
- Arnold, S. E., Franz, B. R., & Trojanowski, J. Q. (1993). Lack of neuropathological findings in elderly patients with schizophrenia. *Neuroscience Abstracts*, 19, 349–350.
- Arnold, S. E., Franz, B. R., & Trojanowski, J. Q. (1994). Elderly patients with schizophrenia exhibit infrequent neurodegenerative lesions. *Neurobiology of Aging*, 15, 299–303.
- Arriagada, P. V., Marzloff, K., & Hyman, B. T. (1992). Distribution of Alzheimer-type pathologic changes in non-demented elderly individuals matches the patterns in Alzheimer's disease. *Neurology*, 42, 1681–1688.
- Berg, L. (1984). Clinical Dementia Rating (letter). *British Journal Psychiatry*, 145, 339.
- Bieliauskas, L. A., & Drag, L. L. (2013). Differential diagnosis of depression and dementia. In L. Ravdin, & H. Katzen (Eds.), *Handbook on the neuropsychology of aging and dementia. Clinical handbooks in neuropsychology*. New York, NY: Springer.
- Brink, T. L. (1984). Limitations of the GDS in cases of pseudodementia. *Clinical Gerontology*, 2, 60–61.
- Christensen, H., Griffiths, K., Mackinnon, & Jacomb, P. (1997). A qualitative review of cognitive deficits in depression and Alzheimer-type dementia. *Journal of the International Neuropsychological Society*, 3, 631–651.

- Crawford, J. R., Stewart, L. E., & Moore, J. W. (1989). Demonstration of savings on the AVLT and development of a parallel form. *Journal of Clinical and Experimental Neuropsychology, 11*, 975–981.
- Cummings, J. L., & Benson, D. F. (1983). *Dementia: A clinical approach* (2nd ed.). Boston, MA: Butterworth's-Heinemann.
- Dahlman, K.L., Davidson, M., & Harvey, P. (1996). Cognitive functioning in late-life schizophrenia: A comparison of elderly schizophrenic patients with Alzheimer's disease. *Paper presented at The Challenge of the Dementias Conference*, The Lancet, Edinburgh, Scotland.
- Davidson, M., Harvey, P. D., Powchik, P., et al. (1995). Severity of symptoms in geriatric schizophrenic patients. *American Journal of Psychiatry, 152*, 197–207.
- Davidson, M., Harvey, P. D., Welsh, K., Powchik, P., Putnam, K., & Mohs, R. C. (1996). Characterization of the cognitive impairment of old-age schizophrenia: A comparison to patients with Alzheimer's disease. *American Journal of Psychiatry, 153*, 1274–1279.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California verbal learning test manual*. San Antonio, TX: The Psychological Corporation.
- Desai, A. K. (2010). Schizophrenia in older adults. *Current Psychiatry, 9*(9), 23A.
- Evans, D. A., Funkenstein, H. H., Albert, M. S., Scherr, P. A., Cook, N. R., Chown, M. J., ... Taylor, J. O. (1989). Prevalence of Alzheimer's disease in a community population of older persons. *Journal of the American Medical Association, 262*, 2551–2556.
- Fromm, D., Holland, A. L., Nebes, R. D., & Oakley, M. A. (1991). A longitudinal study of word-reading ability in Alzheimer's disease: Evidence from the National Adult Reading Test. *Cortex, 27*, 367–376.
- Gatz, M., & Hurwitz, M. L. (1990). Are old people more depressed? Cross-sectional data on center for epidemiological studies depression scale factors. *Psychology of Aging, 5*, 284–290.
- Gold, J. M., & Harvey, P. D. (1993). Cognitive deficits in schizophrenia. *Psychiatric Clinics of North America, 16*(2), 295–312.
- Green, C. R., & Davis, K. L. (1993). Clinical assessment of Alzheimer's-type dementia and related disorders. *Human Psychopharmacology, 4*, 53–71.
- Harding, C. M., Brooks, G. W., Ashikaga, T., Stauss, J. S., & Breier, A. (1987). The Vermont longitudinal study of persons with severe mental illness: II. Long term outcome of subjects who retrospectively met DSM-III criteria for schizophrenia. *American Journal of Psychiatry, 144*, 727–735.
- Harvey, P. D., Lombardi, J. L., Leibman, M., White, L., Parrella, M., Powchik, P., Mohs, R. C., & Davidson, M. (1996). Performance of chronic schizophrenic patients on cognitive neuropsychological measures sensitive to dementia. *International Journal of Geriatric Psychiatry, 11*, 621–627.
- Harvey, P. D., Lombardi, J., Leibman, M., Parella, M., White, L., Powchik, P., ... Davidson, M. (1997). Verbal fluency deficits in geriatric and nongeriatric chronic schizophrenic patients. *Journal of Neuropsychiatry and Clinical Neurosciences, 9*(4), 584–590.
- Harvey, P. D., Powchik, P., Mohs, R. C., & Davidson, M. (1995). Memory functions in geriatric chronic schizophrenic patients: A neuropsychological study. *Journal of Neuropsychiatry and Clinical Neurosciences, 7*(2), 207–212.
- Harvey, P. D., White, L., Parrella, M., Putnam, K. M., Kincaid, M. M., Powchik, P., ... Davidson, M. (1995). The longitudinal stability of cognitive impairment in schizophrenia. Mini-mental state score at one- and two-year follow-ups in geriatric in-patients. *British Journal of Psychiatry, 166*(5), 630–633.
- Heaton, R. K. (1981). *Wisconsin card sorting test manual*. Odessa, FL: Psychological Assessment Resources.

- Heaton, R.K., Chelune, G.J., Talley, J.L., Kay, G.G., & Curtiss, G. (1993). *Wisconsin card sorting test manual* (revised and expanded). Odessa, FL: Psychological Assessment Resources.
- Heaton, R. K., Paulsen, J. S., McAdams, L. A., Kuck, J., Zisook, S., Braff, D., ... Jeste, D. V. (1994). Neuropsychological deficits in schizophrenics: Relationship to age, chronicity, and dementia. *Archives of General Psychiatry*, 51, 469–476.
- Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology of Aging*, 5, 215–227.
- Hooper, H. (1983). *Hooper visual organization test*. Los Angeles: Western Psychological Services.
- Hotchkiss, A. P., & Harvey, P. D. (1990). Effects of distraction on communication failures in schizophrenic patients. *American Journal of Psychiatry*, 4, 513–515.
- Huff, F. J., Growdon, J. H., Corkin, S., & Rosen, T. R. (1987). Age at onset and rate of progression of Alzheimer's disease. *Journal of American Geriatric Society*, 35, 27–30.
- Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992). Mayo's older americans normative studies: WAIS-R norms for ages 56–97. *Clinical Neuropsychologist*, 6(Suppl), 1–30.
- Ivnik, R. J., Smith, G. E., Malec, J. F., Petersen, R. C., & Tangalos, E. G. (1995). Long-term stability and inter-correlations of cognitive abilities in older persons. *Psychological Assessment*, 7, 155–161.
- Jastak, S., & Wilkinson, G. (1984). *The wide range achievement test-revised*. Wilmington, DE: Jastak Associates.
- Jeste, D. V. (1993). Late life schizophrenia: Editor's introduction. *Schizophrenia Bulletin*, 19, 687–689.
- Jorm, A. F., & Jacomb, P. A. (1989). The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): Socio-demographic correlates, reliability, validity and some norms. *Psychological Medicine*, 19(4), 1015–1022.
- Jorm, A. F., Scott, R., & Jacomb, P. A. (1989). Assessment of cognitive decline in dementia by informant questionnaire. *International Journal of Geriatric Psychiatry*, 4(1), 35–39.
- Juva, K., Sulkava, R., Erkinjuntti, T., Ylikoski, R., Valvanne, J., & Tilvis, R. (1994). Staging the severity of dementia: Comparison of clinical (CDR, DSM-III-R), functional (ADL, IADL) and cognitive (MMSE) scales. *Acta Neurologica Scandinavica*, 90, 293–298.
- Kasznaik, A. W., & Christensen, G. D. (1994). Differential diagnosis of dementia and depression. In M. Storandt, & G. R. Vandenberg (Eds.), *Neuropsychological assessment of dementia and depression in older adults: A clinician's guide*. Washington, DC: American Psychological Association.
- Katzman, R., Brown, T., Fuld, P., et al. (1983). Validation of a short orientation–memory–concentration test of cognitive impairment. *American Journal of Psychiatry*, 140, 734–739.
- Kincaid, M. M., Harvey, P. D., Parrella, M., White, L., Putnam, K. M., Powchik, P., ... Mohs, R. C. (1995). Validity and utility of the ADAS-L for measurement of cognitive and functional impairment in geriatric schizophrenic inpatients. *Journal of Neuropsychiatry and Clinical Neurosciences*, 7, 76–81.
- King, D. A., Cox, C., Lyness, J. M., Conwell, Y., & Caine, E. D. (1998). Quantitative qualitative differences in the verbal learning performance of elderly depressives and healthy controls. *Journal of International Neuropsychological Society*, 4, 115–126.
- Kumar, A., & Gottlieb, G. (1993). Frontotemporal dementias. *American Journal of Geriatric Psychiatry*, 1, 95–107.

- Lafleche, G., & Albert, M. (1995). Executive functioning deficits in mild Alzheimer's disease. *Neuropsychology, 9*, 313–320.
- Mattis, S. (1976). Mental status examination for organic mental syndrome in the elderly patient. In L. Bellak, & T. B. Karasu (Eds.), *Geriatric psychiatry*. New York: Grune & Stratton.
- Mohs, R. C., Rosen, W. G., Greenwald, B. S., & Davis, K. L. (1983). Neuropathologically validated scales for Alzheimer's disease. In T. Crook, S. Ferris, & R. Bartus (Eds.), *Geriatric psychopharmacology* (pp. 37–45). New Canaan, CT: Mark Powley Associates.
- Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology, 43*, 2412–2414.
- Morris, J. C., McKeel, D. W., Storandt, M., Rubin, E. H., Price, J. L., Grant, E. A., ... Berg, L. (1991). Very mild Alzheimer's disease: Informant-based clinical, psychometric, and pathologic distinction from normal aging. *Neurology, 41*, 469–478.
- Paveza, G. J., Cohen, D., Eindsdorfer, C., et al. (1992). Severe family violence and Alzheimer's disease: Prevalence and risk factors. *Gerontologist, 32*(4), 493–497.
- Powchik, P., Davidson, M., Nemeroff, C. B., Haroutunian, V., Purohit, D., Losonczy, M., ... Davis, K. L. (1993). Alzheimer's disease related protein in geriatric, cognitively impaired schizophrenic patients. *American Journal of Psychiatry, 50*, 1726–1727.
- Prohovnik, I., Dwork, A. J., Kaufman, M. A., & Wilson, N. (1993). Alzheimer's type neuropathology in elderly schizophrenia. *Schizophrenia Bulletin, 19*, 805–816.
- Purohit, D. P., Davidson, M., Perl, D. P., Powchik, P., Haroutunian, V. H., Bierer, L. M., McCrystal, J., Losonczy, M., & Davis, K. L. (1993). Severe cognitive impairments in elderly schizophrenic patients: Clinicopathologic study. *Biological Psychiatry, 33*, 255–260.
- Putnam, K. M., Harvey, P. D., Parrella, M., White, L., Kincaid, M., Powchik, P., & Davidson, M. (1996). Symptom stability in geriatric chronic schizophrenic inpatients: A one-year follow-up study. *Society of Biological Psychiatry, 39*, 92–99.
- Rebok, G. W., & Folstein, M. F. (1993). Dementia. *Journal of Neuropsychiatry and Clinical Neurosciences, 5*, 265–276.
- Richie, K., Ledesert, B., & Touchon, J. (1993). The Eugenia study of cognitive ageing: Who are the 'normal' elderly? *International Journal of Geriatric Psychiatry, 8*, 969–977.
- Silverman, J. M., Breitner, J. C. S., Mohs, R. C., & Davis, K. L. (1986). Reliability of the family history method in genetic studies of Alzheimer's disease and related dementia. *American Journal of Psychiatry, 143*, 1279–1282.
- Snowdon, J. (1990). Validity of the Geriatric Depression Scale. *Journal of the American Geriatrics Society, 38*, 722–723.
- Snowdon, J., & Donnelly, N. (1986). A study of depression in nursing homes. *Journal of Psychiatric Research, 20*, 327–333.
- Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary* (2nd ed.). New York: Oxford University Press.
- Stebbins, G. T., Gilley, D. W., Wilson, R. S., et al. (1990). Effects of language disturbances on premorbid estimates of IQ in mild dementia. *The Clinical Neuropsychologist, 4*, 64–68.
- Stebbins, G. T., Wilson, R. S., Gilley, D. W., et al. (1990). Use of the National Adult Reading Test to estimate premorbid IQ in dementia. *The Clinical Neuropsychologist, 4*, 18–24.
- Teri, L., & Wagner, A. (1992). Alzheimer's disease and depression. *Journal of Consulting and Clinical Psychology, 60*, 379–391.

- Wechsler, D. (1987). *Wechsler adult intelligence scale manual* (revised ed.). San Antonio, TX: The Psychological Corporation.
- World Health Organization, (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.
- Yesavage, J. (1987). The use of self-rating depression scales in the elderly. In L. W. Poon (Ed.), *Handbook for clinical memory assessment of older adults*. Washington, DC: American Psychological Association.
- Zubenko, G. S., Sullivan, P., Nelson, J. P., et al. (1990). Brain imaging abnormalities in mental disorders of late life. *Archives of Neurology*, 47, 1107–1111.

# Forensic psychology: practice issues

18

Arthur MacNeill Horton, Jr.<sup>1</sup> and Henry V. Soper<sup>2</sup>

<sup>1</sup>Psych Associates of Maryland, Towson, Columbia, MD, United States, <sup>2</sup>Fielding Graduate University, Santa Barbara, CA, United States

The history of psychology in the legal arena is relatively brief. Freud (1906) was one of the first clinicians to suggest that the study of mental processes would be of use in the legal arena. Early psychologists to suggest a forensic role for psychology were Munsterberg (1908) and Watson (1913). More recent comments note further development of forensic psychology (Blau, 1998; Kurke, 1980; Otto & Heilbrun, 2002). The practice of forensic psychology is quite different in many ways than the practice of other types of psychology (e.g., clinical, counseling, school, etc.). In this chapter, a number of topics related to forensic psychology will be discussed. These will include the question of what forensic psychology is and how it differs from other types of psychological practice. Also covered will be the role of a forensic expert, forensic methods, and validity of data (Block & Berman, 2015). Communications with parties such as patient, attorneys and the court are complex in forensic psychology and include topics such as informed consent, record keeping, and responding to subpoenas. Often forensic psychology practice will require courtroom testimony and there are various standards for testifying, rules of evidence including the basis of opinions and dealing with the ultimate issue.

## What is forensic psychology?

Simply put forensic psychology “...refers to professional practice by any psychologist working within any subdiscipline 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” p. 7 (APA, 2013a).

In other words, forensic psychology is where psychology and the law intersect and the methods, theories, and concepts of psychology can inform the court in helping the court to answer legal questions. Note that the court asks and ultimately answers the legal questions and psychology assists in providing a basis for deciding the correct answers to the legal questions (Blau, 1998). As can be seen, forensic psychology does not depend on the specific area of psychological practice as legal questions can overlap with multiple areas of psychology. Put another way, the forensic role is determined by the service provided and the fact the service is

provided in a legal context or environment which may or may not be within a courtroom (Block & Berman, 2015). Ultimately, forensic psychology provides an opinion or opinions regarding an issue important in a legal matter. The opinions must be based on “adequate scientific foundation, and reliable and valid principles and methods that have been applied appropriately to the facts of the case.” p. 9 (APA, 2013a).

Forensic psychology is a fast-growing subspecialty of psychology, where the results of assessment are used to address legal questions within the courtroom or other such legal arenas (Horton & Hartlage, 2003). However, when psychologists enter the forensic arena there are several challenges to overcome. There are differences between traditional clinical psychological assessments and forensic evaluations. These include the need to relate the psychological findings to the legal issues of the forensic setting, the greater need for symptom validity and effort testing, as well as communicating psychological implications for legal issues in ways that judges and juries without scientific backgrounds can comprehend so that they can make legal decisions (Heilbrun et al., 2003). In the forensic setting, the psychological report must deal with the adversarial nature of legal proceedings and the implications. In the legal setting, the issues to be decided are fought over until there is a winner, and lawyers will treat the forensic report as a weapon (Zillmer & Green, 2006). This is different from the clinical setting, in which a clinical report is treated with great respect by colleagues. Nonetheless, the forensic report must address the implications of the psychological findings for the legal questions at hand in a direct matter, yet not attempt to answer the ultimate legal question, since that is the responsibility of triers of fact, the judge and jury (Blau, 1998).

Since some would have much to gain if the legal issues were decided in their favor, there is a great need for symptom validity testing. Effort testing arises from the fact that potential for gain, monetary and otherwise, may influence the performance on the psychological tests (Iverson, 2003). The likelihood is that a greater number of people seen in the forensic setting will give poor effort than in clinical settings (Mittenberg, Patton, Canyock, & Condit, 2002). Therefore, the forensic psychologist must carefully consider the possibility of the patient giving poor effort and include symptom validity testing as an essential portion of the forensic psychological evaluation (Reynolds, 1998).

The purpose of the forensic psychology report is to educate the readers (i.e., lawyers, judges, and jury members) about the implications of the assessment findings for the legal questions at hand. The forensic report should be so clear that the least informed jury member can understand the conceptual reasoning, and it should provide vivid examples of the legal implications of the forensic evaluation (Heilbrun et al., 2003).

## **Legal courts system**

Most often forensic psychology is practiced in courts. Perhaps the three major types of courts are family courts, civil courts, and criminal courts. In family courts, the

focus is on the family with some jurisdictions having a dedicated court with specially assigned judges who only decide family issues. Family issues frequently occur following divorce; for example, child custody evaluations involving parental conflicts regarding children with visitation risk assessment, grandparent visitation assessment, and termination of parental rights assessments (Weithorn, 2006; 1984). In addition, child abuse and adoption readiness evaluations may be requested.

Civil courts deal with disputes between different parties that involve money. For example, personal injury evaluations may assess the degree of damages following an accident or other wrongful acts such as medical malpractice. Typically a plaintiff claims an injury for which a defendant is liable for monetary damage (Hartlage, 2003). The defendant may request an Independent Medical Examination (IME) Second Opinion by a forensic psychologist to dispute the claim of injury. Personal injuries on the job may be handled by Workman's Compensation Evaluations and if a person is thought not competent to handle their legal/fiscal/medical treatment affairs, then a civil competency evaluation may be done to determine if a guardianship arrangement is in the person's best interest (Franzen, 2010).

In contrast, criminal courts deal with persons accused by the state of breaking the law. Evaluations may be either before it has been decided the person is guilty of breaking the law or after the court has decided the person is guilty in deciding the type and extent of punishment. Prior to deciding guilt or innocence, forensic psychology evaluations may be done to determine the person's ability to stand trial and assist their lawyers, or their mental state at the time of the crime (Denney, Tyner, & Hartlage, 2010). After the trial, the question of mitigating factors becomes important in determining the punishment for the crime. For example, being mentally retarded (e.g., term now changed to intellectual disability) is a legal basis for not being put to death, and intelligence testing is a key piece of forensic psychology information in determining mental retardation.

## Forensic roles

In addition to courtroom testimony as an expert witness, there are other forensic roles a forensic psychologist may play (APA, 2013a). These may be either advisory/consultative or experimental. For example, in an advisory/consultative role, a forensic psychologist may advise lawyers regarding psychological research literature and the psychological evidence of other experts. In addition, a forensic psychologist may assist in preparing the cross examination of other witnesses. In contrast, the experimental role deals with psychological knowledge and jury selection, problems with eyewitness identification and using children in the courtroom as witnesses (Weithorn, 1984).

The role of the expert witness is perhaps the role most associated with the forensic psychologist (Blau, 1998). It is most important to first differentiate expert witness from fact witnesses. Simply put, fact witnesses are persons without recognized professional expertise who are eye witnesses to aspects of the legal matter the court

is attempting to decide. In other words, they personally saw and/or heard information important to deciding the legal question (Block & Berman, 2015); for example, a person who witnessed a motor vehicle accident.

In contrast, an expert witness is a person with recognized professional expertise who can testify regarding an opinion or opinions based on scientific research that can assist the court in deciding a legal question (Blau, 1998). Forensic psychologists who are requested to psychologically evaluate a person are considered expert witnesses because their expertise is why they are involved in the case (Block & Berman, 2015).

Expert witnesses assist the court in resolving legal matters by providing opinions not available to a person without the recognized professional expertise of the forensic psychologist. In courtroom testimony, persons called as expert witness go through Voir Dire, which is a description of the expert witnesses' education, training, and professional experience and scientific contributions which support the expert witnesses' expert status, in response to which the judge makes a determination in court if a witness is considered an expert witness and then allowed to testify regarding the legal case (Blau, 1998).

Expert witnesses may testify for either the defendant or plaintiff side, or in some cases are retained by the court rather than either side to provide testimony. Expert witnesses are expected to not be advocates for either the defendant or plaintiff side but rather to be unbiased (APA, 2013a). The functions of the expert witness are to educate the judge and jury regarding the psychological aspects of the legal matter to be decided and provide an opinion. While the individual court judge always qualifies a witness as expert based on the witness' education, training and experience, rules for expert witnesses vary by which State or Federal Court in which the case is being tried. For example, Federal Rules of Evidence only apply to Federal Courts (Crown, Fingerhut & Lowenthal, 2010).

Also, expert witnesses should be careful to identify their areas of expertise and acknowledge their areas of uncertainty. For forensic psychologists as expert witnesses, it is important to be aware of what the specific court rules are regarding testimony by psychologists. For example, in some states the person's diagnosis is considered a medical question only to be answered by a physician.

## **U.S. legal decisions on scientific expertise**

Courts have set standards for judging the value of expert witness testimony. The older standard from *Frye v. United States* (1923) focused on the general acceptance of the scientific principle or discovery. In other words, did most scientists in the field accept the science supporting the testimony? There had been some concerns that the Frye standard had become too lax and that "junk science" testimony was invading the courts (Project on Scientific knowledge and Public Policy, 2003). In response, the *Daubert v. Merrell Dow Pharmaceuticals* standard (*Daubert v. Merrell Dow*, 509 US 579, 1993) was established.

This decision set up new rules for the evaluation of scientific evidence with an emphasis on the validity of the science supporting the testimony and publication of the science in peer-reviewed journals. The legal standard under Frye test was the evidence must be a result of a theory that has “general acceptance” in the scientific community ([Frye v. United States, 1923](#)), in other words, a popular theory or finding. In contrast, the Daubert test was more detailed and objective ([Project on Scientific knowledge and Public Policy, 2003](#)). Under Daubert the evidence need to satisfy four criteria:

1. The theory is testable.
2. The theory has been peer-reviewed.
3. The reliability and error rate must be known.
4. The theory must be generally accepted in the scientific community.

In other words, the Daubert test included the Frye test but added additional requirements. It might be mentioned that where the Daubert criteria say “reliable” that is the legal phrasing and what is really intended is the theory is considered “valid” or true. So for the Daubert test in addition to the Frye test the testimony needed to also have been peer-reviewed, testable, and have a recognized error rate (sensitivity and specificity data). While most states have adopted the Daubert test, eight states still retain the Frye test ([Block & Berman, 2015](#)).

## Forensic process

In forensic psychology evaluations, it is recommended that a forensic assessment checklist of topic heading be used to ensure all relevant issues and matters have been considered ([Block & Berman, 2015; Grisso, 2010](#)).

First, referral information is documented. The referral source, reason for the evaluation, questions to be answered and case name, court, and docket should all be stated.

Second, the applicable standards and laws should be specified. These may vary by jurisdiction or type of court but will likely be important in framing opinions.

Third, documents reviewed need to be listed. Each document needs to be identified, including its source.

Fourth, collateral contacts need to be listed by dates. For example, each person interviewed and/or providing records.

Fifth, informed consent for forensic psychology assessment must be documented in writing. All informed consent documents should include the nature (nontherapeutic relationship with a forensic psychologist), purpose (e.g., legal questions to be addressed), and procedures of the evaluation (e.g., interviews, observations, psychological testing, interviews, and ratings of third-party collaterals, or record reviews), possibility the findings of the forensic psychology evaluation may or may not be helpful or emotionally distressed to the patient, who has employed the forensic psychologist, confidentiality included limits, who will have access to data (including mandated reporting of sexual abuse allegations and communications with attorneys), feedback to be provided, intended use of evaluation findings, and responsibility for payment.

Sixth, contextual issues should be mentioned such as the psychological and social ecology, that is, divorce or criminal charges, etc.

Seventh, third-party involvement, obligations, or entitlements should be specified. For example, who will have access to forensic psychology findings other than the person assessed, and who will be responsible for payment.

Eighth, current status observations should be provided. These are similar to other psychology reports where behaviors observed during the interview such as affect, mood, attention and concentration, form and flow of thought processes, thought content, perceptual processes (e.g., absence of hallucinations and delusions), speech, and motor abilities are mentioned.

Ninth, any deviations from standard practice needed to accommodate the patient should be noted. For example, third-party observer effects have been well documented in the psychology research literature. If there were any observers present during psychological testing or if the psychological testing was recorded by audiovisual means, this information should be documented and a rationale for the deviation from standard practice provided.

Tenth, a complete list of all psychological tests administered should be provided. Descriptive information and references should be included for any unusual instruments or techniques used ([Archer, Buffington-Vollum, Stredny, & Handel, 2006](#); [Hartlage, 2010](#); [Lally, 2003](#)).

Eleventh, reliability and validity of the psychological test results should be mentioned ([Melton, Petrilà, Poythress, & Slobogin, 2007](#); [Reynolds & Livingston, 2012](#); [Strauss, Sherman, & Spreen, 2006](#)). Recent psychology research has developed specific standalone tests to evaluate the degree of effort provided on psychological testing ([Iverson, 2010](#)). Current practice is to include at least two specific tests of effort in every forensic psychology evaluation ([Bush et al., 2005](#)). At the same time, it would be important to comment on any mediating factors, such as dementia, which could have an effect on the results of effort testing.

Twelfth, data presentation of the forensic psychological test results needs to explain the results of all of the psychological tests administered. The meaning of the data discussed should be in the context of the legal referral questions that prompted the forensic psychology assessment ([Block & Berman, 2015](#)).

Thirteenth, a summary that includes the most important aspects of the earlier portions of the forensic psychology report should be presented. The summary should not mention material not previously addressed in the body of the report ([Block & Berman, 2015](#)).

Fourteenth, any recommendations made should logically flow from the summary and should be related to the initial legal referral questions. Diagnoses may or may not be needed depending on the legal referral question. In some states, psychologists are not allowed to make medical diagnoses so any conclusions of a diagnostic nature are termed “diagnostic impressions.” Diagnoses can either be based on the current Diagnostic and Statistical Manual (DSM) of the American Psychiatric Association (AMA) ([American Psychiatric Association, 2013](#)) or the current International Classification of Diseases (ICD) of the World Health Organization ([World Health Organization, 2005](#)). If diagnoses are given then it should be

specified whether they are DSM or ICD. DSM-5 diagnoses may have new implications for the practice of forensic psychology (Hopwood & Sellbom, 2013). In addition, all forensic psychology reports should be signed by the person or persons who conducted the forensic psychology evaluation and include the academic degrees and professional title of persons signing (Block & Berman, 2015).

## General procedures—collateral sources

Collecting information from collateral sources (third-party individuals and/or records) is a mandatory component of most forensic psychology evaluations. This is because traditional methods of psychological assessment (e.g., interviewing and psychological testing) have inherent limitations as means of obtaining accurate information in the forensic arena. Collateral sources of information may include police and criminal history records, medical records, employment records, educational records including standardized test scores and grades and interviews and rating scales filled out by significant others, and more recently using the Internet and social media to collect information (Pirelli, Otto, & Estoup, 2016). Significant others may include family members, personal friends, co-workers, administrative supervisors, treating physicians, and therapists.

Collateral sources must be informed that the information is to be recorded in notes and maintained and there is no confidentiality as the information is being obtained in the context of litigation. As noted by the APA (2013a) (p. 14), “Forensic practitioners strive to access information or records from collateral sources with the consent of the relevant attorney or the relevant party, or when otherwise authorized by law or court order.”

Collateral sources are often used to confirm or reject hypotheses and are most useful when the collateral is low on affiliation with a person in litigation but has significant information (Block & Berman, 2015). Collateral sources are most important in reconstructive forensic psychology evaluations (e.g., insanity, testamentary capacity, and assessing pre-existing psychological conditions) (Gottfried, Schenk, & Vitacco, 2016).

## Assessment tools

Psychological testing must be reliable and valid for the purpose used (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). A common misunderstanding regarding psychological testing is that psychological testing is about individual psychological test scores. Rather, psychological testing is the interpretation of psychological test scores by a professional psychologist (Reynolds & Livingston, 2012). The key concept is interpretation by a professional psychologist, as the interpretation of the psychological test scores includes consideration of the relevant

psychological research literature regarding the psychological test that was used to obtain the psychological scores by a professional psychologist (Reynolds & Livingston, 2012).

The interpretation is important because the sample of persons the psychological test was normed on may differ from the sample of persons to which the interpretation of the psychological test scores is being applied. Different samples may require different interpretations of psychological test scores as the psychology research literature may have found sample based differences (Reynolds & Livingston, 2012). There should be empirical support (peer-reviewed research publications) for the intended use of the psychological tests. Also, there have been challenges to the use of certain psychological tests (e.g., Rorschach) for forensic purposes (Block & Berman, 2015).

It is important to be aware of the research findings regarding any psychological tests used in a forensic psychology evaluation. As previously noted the Frye and Daubert legal decisions have criteria that would be applied to the admissibility of psychological testing.

In general, traditional clinical psychological tests can provide an understanding of an individual's intellectual, emotional, and personality functioning and have extensive research bases and well-established psychometric properties (Reynolds & Livingston, 2012).

On the other hand, the original development of traditional clinical psychological tests focused on diagnosis and treatment planning rather than forensic applications, and in addition, few clinical psychological tests have appropriate forensic norms and therefore results from traditional clinical psychological tests have limitations when used to address specific legal issues (Block & Berman, 2015).

In some cases, specialized psychological tests have been developed which have a focus on legal issues. For example, psychological tests of effort are often used to assess the probability that an individual has been malingering in a legal context (Iverson, 2010; Larrabee, 2005; Reynolds & Horton, 2012; Rogers, 2008) and there has been extensive normative data on forensic populations (Mittenberg et al., 2002). Unfortunately, there are numerous legal issues for which there are no specialized forensic psychological tests. The use of psychometrists administering psychological tests under the supervision of a licensed doctoral level psychologist is well established in clinical psychology practice and appropriate in forensic cases (Malek-Ahmadi, Erickson, Puente, Pliskin, & Rock, 2012).

## Forensic role function

It is important to realize that, in forensic psychology, there are many role differences relative to the traditional psychotherapy role. For example, in forensic psychology evaluations, the person/agency being served is an attorney or the court with resolving a legal issue and there is not a traditional therapeutic relationship with the person being assessed but rather there may be an adversary relationship

depending on legal circumstances (Block & Berman, 2015). Ethical decision making (Bush, Connell, & Denney, 2006) may be much more complex than in more traditional areas of psychological practice.

## Courtroom testimony

The classic references for psychologists preparing for courtroom testimony are two excellent books by Stanley L. Brodsky, on testifying in Court (2013) and being an expert witness (2016) published by the American Psychological Association (APA). Brodsky provides excellent practical advice for successful courtroom testimony and are required reading for every forensic psychologist. Also see Gutheil (2009), Gutheil and Dattilio (2007), Hartlage (2003), Hartlage & Stern (2010), Malone and Zweier (2014), Otto, DeMint, and Boccaccini (2014), Weiss and Watson (2015) for additional practical advice.

Often prior to courtroom testimony, there is a deposition prior to a legal case going to trial. At the deposition, there is a court reporter taking down the forensic psychologist's testimony during questioning by both attorneys involved in the case (Hartlage & Williams, 2010). The purpose of the deposition is for the opposing side to discover the testimony that the forensic psychologist is to offer in a trial (Blau, 1998). Everything said about courtroom testimony is applicable to deposition testimony. Often, legal cases may settle after a deposition and never go to trial, but the deposition may be brought up at trial and courtroom testimony should not be inconsistent with deposition testimony.

For testifying in court there is a specific process when the forensic psychologist is an expert witness. First there is Voir Dire which, as earlier explained, is when the attorney who retained the forensic psychologist questions the forensic psychologist regarding the forensic psychologist's education, training, professional experiences and scientific contributions (Blau, 1998). Table 18.1 presents a script of attorney questions for Voir Dire.

The other side's attorney then has the opportunity to question the forensic psychologist further and then the judge decides whether to accept the forensic

**Table 18.1** Courtroom testimony script—Voir Dire

Please state your full name, office address for this case and occupation or profession.
Please tell the jury your educational background.
Please tell the jury in which states you're licensed to practice psychology.
Please tell the jury what have been your professional experiences.
Please tell the jury the professional organizations to which you belong.
Please tell the injury the professional associations of which you've been elected president.
Please tell the jury if you have ever served on a state psychology licensing board.
Please tell the jury what you've published in your field.
Please tell the jury if you're board certified.
Have you ever appeared as expert witness in other legal cases prior to this time?

psychologist as an expert witness. After being accepted as an expert witness, the forensic psychologist offers direct testimony regarding the forensic psychologist's findings in response to questions asked by the retaining attorney. Direct testimony may include information from the clinical interview, behavioral observations, psychological test results, review of medical, educational, legal, and employment records and information obtained from collaterals and the forensic psychologist's interpretation of the multiple sources of information (Brodsky, 2013). Table 18.2 presents a script of attorney questions for direct testimony.

After direct testimony has concluded, there is the opportunity for cross-examination by the opposing attorney. The purpose of cross-examination is to question the validity of the forensic psychologist's direct testimony (Blau, 1998).

The tone of the opposing attorney's questioning is often clearly adversarial during cross-examination. After cross-examination has been completed, the retaining attorney has the opportunity to conduct a redirect questioning. Essentially the redirect is an opportunity for the retaining attorney to clarify any possible misunderstandings that may occur during the cross-examination. After the redirect questioning has concluded however, the opposing attorney has an opportunity to conduct a recross examination to address issues in the redirect questioning. As can be seen courtroom testimony is a very adversarial process where a forensic psychologist's findings and opinions are subjected to hostile scrutiny and skeptical questioning.

When providing courtroom testimony, a forensic psychologist must provide opinions that are clearly supported by the psychological test data and other sources of information that were reviewed. As earlier mentioned, it is important to realize the limits of the forensic psychologist's expertise and not to offer opinions that go beyond that expertise (APA, 2013a). In offering testimony, it is thought to be important to avoid detailed overly technical descriptions and explain forensic psychological constructs and findings in ways that are understandable to a nonpsychologist.

Also it is important to realize that the forensic psychologist as an expert witness is expected to take an independent evaluator's role and not to advocate for either side in a legal dispute. Rather the forensic psychologist is an advocate for the interpretations of these data, not for either party in a legal dispute (APA, 2013a).

**Table 18.2** Direct testimony script

In your practice of psychology have you had occasion to see the plaintiff? Where and when did you see the plaintiff (defendant)? By whom was the plaintiff (defendant) referred to you? For what purpose was the plaintiff (defendant) referred to you? How long did that psychological evaluation take? Please tell the jury what the psychological evaluation consisted of. Please tell the jury what your findings were with regard to the plaintiff (defendant) as a result of your psychological evaluation.
---

## Expert witness fees

Contingency fees (payment depends on winning the legal case) are specifically prohibited by the Specialty Guidelines for Forensic Psychologists (2013). The reason for prohibiting contingency fees is that having a financial interest in the outcome of a legal case (only paid if the case is won) impairs the impartiality of the forensic psychologist. In order to adequately assist the trier of fact in resolving the legal issue, the forensic psychologist must maintain impartiality regarding payment.

As expert witnesses, forensic psychologists charge for their professional time and special psychological expertise, that was acquired through education, training and professional experience. Usually charging is hourly and is similar to attorneys charging by the 0.1 hour (6 minutes) (Block & Berman, 2015). Fee arrangements must be in writing and very specific and agreed to in writing (signed) by both the attorney and forensic psychologist prior to doing any forensic psychology work.

The fee agreement should include clear statements regarding the method by which fees are determined (when and how) and the terms of payment (additional charges for fees not paid within 30 days of date of invoice). The fee agreement should indicate billable time including travel time (portal to portal), preparation time, document review time, consultation with attorney and interview of patient and collaterals, and psychological testing time (Block & Berman, 2015).

The preferred method of handling forensic psychology payment is to have a written fee agreement already signed by the attorney and to request a retainer for the estimated amount of professional time to be used, and not do any work until the retainer is received. Then, bill against the retainer until the retainer is exhausted and then request the retainer be replenished for the estimated remaining future work prior to continuing work. Not all attorneys will agree to advance a retainer, but having been advanced a retainer does prevent doing professional forensic psychology work and accumulating unpaid bills.

## Ethical issues

All forensic psychologists should be very familiar with the American Psychological Association (APA) Specialty Guidelines for Forensic Psychology (APA, 2013a,b,c) and the APA Ethical Code (APA, 2010a,b), and ethical decision making related to the APA Ethical Code (Bush et al., 2006) and their individual State statutory regulations and legal standards for forensic testimony.

In addition, APA has Guidelines for assessment of older adults with diminished capacity, multicultural education, training, research, practice, and organizational change for psychologists (2003), psychological practice with older adults, (2004), record keeping guidelines (2007), child custody evaluations in family law proceedings (2010), psychological evaluation in child protection matters (2013), practice of parent coordination (2012), evaluation of dementia and age-related cognitive change (2012), assessment of and intervention with persons with disabilities (2013),

and psychological practice with lesbian, gay, and bisexual clients (2012), and it would be important for a forensic psychologist to be familiar with any relevant guidelines. As the legal maxim says, “Ignorance of the law is no exception or reason for not following the law.”

## Multiple relationships

As noted by the APA Specialty Guidelines for Forensic Psychology, “A multiple relationship occurs when a forensic practitioner is in a professional role with a person and, at the same time or at a subsequent time, is in a different role with the same person; is involved in a personal, fiscal, or other relationship with an adverse party; at the same time is in a relation with a person closely associated with or related to the person with whom the forensic practitioner has the professional relationship; or offers or agrees to enter into another relationship in the future with the person or a person closely associated with are related to the person (EPPC Standard 3.05)” (p. 11, [APA, 2010a,b](#)). Multiple relationships may include therapist to evaluator, evaluator to therapist, and therapist to expert witness, among others. Obviously, a forensic psychologist does not provide services to a family member or anyone with whom they have a close personal or fiscal relationship as the prior and/or current relationship may compromise the objectivity of the forensic psychologist and/or do harm to others.

When requested to perform concurrent or sequential forensic or therapeutic services, the forensic psychologist discloses the potential for risk and impairment of objectivity to all involved in the legal matter and makes reasonable efforts to refer the request to another qualified forensic psychologist, if one is available ([APA, 2013a](#)). In some cases, a court will order a forensic psychologist to perform concurrent and/or sequential services and the court may not be willing to change their order despite any concerns on the part of the forensic psychologist ([Crown et al., 2010](#)).

Sometimes courts feel that former treating doctor psychologists are preferred as expert witnesses to a forensic psychologist who entered the case only at the request of the attorney on one side of the case. There are forensic psychologists who feel that forensic psychologists should never serve as a therapist and an evaluator ([Greenberg & Shulman, 1997](#); [Greenberg & Shulman, 2007](#)), but there is not a complete ban on the practice in the [APA \(2013a\)](#).

## Working within the legal system

Always return telephone calls from attorneys within 24 hours (return all telephone calls from a judge immediately!!) even to say you can't discuss the matter until you have a signed release from your patient or that you do not know the person ([Block & Berman, 2015](#)). As mentioned earlier, it is very important to be familiar with legal procedures and statutes, case law, and rules of evidence before beginning

forensic psychological work, and remember lawyers get paid at the beginning of the process and forensic psychologists should also insist on getting paid at the beginning. Lawyers and psychologists have very different ethical guidelines and rules of conduct—be aware of the differences (Foote, 2006).

Beware of attempts by an attorney to expand your testimony into areas beyond your competence and the legal statutes in your jurisdiction for a forensic psychologist. In communications with attorneys, if serving as an expert witness for one side, only communicate with that attorney, despite attempts by the other side's attorney to communicate with you. If an expert witness is appointed by the court, both attorneys should be present or on a conference call. All sources of information should be disclosed during discovery. Regarding the work of other experts, it is important to focus critical comments on the data, opinions, and recommendations of the other expert and to maintain personal respect for the other expert (Block & Berman, 2015).

## Documentation and record keeping

In forensic psychology practice, the American Psychological Association (2007) has published Guidelines for record keeping, and many states have state laws that give standards and requirements for record keeping and document retention and release, and federal agencies (e.g., Medicare, Social Security Administration, etc.) and managed care organizations have their own requirements (Block & Berman, 2015).

It is important to be aware of guidelines, laws, and regulations regarding record keeping and releasing records that apply in the jurisdiction a forensic psychologist practices (Bush & Martin, 2010). How long records are required to be retained may vary depending on jurisdiction and/or the organizations requiring the record keeping. In general, it is important to document all contacts and interviews with patients and collaterals, attorneys, and any other interested parties such as the court. All testing/evaluation raw data (including video and/or audio recordings), notes and correspondence (including letters, emails, texts, and letters) scoring and reports, billing records and medical, educational, psychological, employment, and law enforcement and legal records reviewed need to be documented and retained.

The court needs to be able to reconstruct all data that lead to an expert witness' conclusions and opinions (Block & Berman, 2015). In the discovery phase prior to trial, records will need to be disclosed to the opposing side. Raw psychological test data is protected in some jurisdictions and can only be released to another qualified psychologist, but this varies by jurisdictions and it is important to know what applies in the jurisdiction in which a forensic psychologist practices (Bush, Rapp & Ferber, 2010).

Forensic psychology records should demonstrate a consistency of data leading to findings—which in turn lead to opinions—so that it will be clear how opinions were reached. The informed consent form signed prior to evaluation should include statements regarding record keeping and record release. It is important to document release of records and a release of information form should be written; it should

include the name of the forensic psychologist releasing the records, the name of the person to whom the records are to be released, and dated and signed by the “person in interest.” The release of information form should also identify what records are to be released and state the period of time for which the form will be valid (Block & Berman, 2015).

## Dealing with subpoenas

In brief, a subpoena is a formal request from an attorney to require a witness to send the attorney records and/or summon an expert witness to a deposition and/or trial to give testimony and bring documents (Blau, 1998). While not a court order, which comes from a judge, subpoenas require a response and cannot be ignored by a forensic psychologist.

The recommended approach is to telephone the attorney who sent the subpoena to clarify the purpose of the subpoena and the information being requested and to document the telephone conversation in writing but not to immediately disclose the requested information. Rather consult your own attorney as to how to respond to the requesting attorney (Block & Berman, 2015). A useful legal maxim is “Whoever serves as their own attorney has a fool for a client.” The subpoena must have been properly served (in person or to your employee) to be valid, it must give proper notice in terms of time (number of days prior to deposition, trial and document release), and in some cases the request may have been unreasonable (i.e., requesting an individual’s mental health records without providing a copy of the documented informed consent of the individual), but these are all legal questions to be answered by the forensic psychologist’s attorney (Block & Berman, 2015).

For example, you are an expert witness and the attorney sending the subpoena claims you are a fact witness and the attorney does not need to pay your expert witness fees for your professional time in a courtroom and/or a deposition. In such cases, the forensic psychologist needs to retain their own attorney to communicate with the other attorney and inform the other attorney that the forensic psychologist is an expert witness. A letter from the forensic psychologist’s attorney can also communicate that the forensic psychologist is an expert witness and that professional fees need to be paid in advance of any appearance at a deposition and/or trial.

Also the forensic psychologist’s attorney can, if the attorney to attorney communication does not work, file a motion to quash the subpoena in court. A motion to quash goes before a judge and requires a legal rationale. Treatment records, for example, are the property of the individual who was treated, and that person must consent in writing to the release of the records before the records can be released (Bush et al., 2010).

Depending on the jurisdiction, fees may be charged for release of records for photocopying, handling, and mailing charges. These may vary by jurisdiction and state laws and other federal laws and regulations will need to be consulted prior to charging fees.

## Summary

Forensic psychology is a new and emerging area of psychological practice. There is much that psychology can offer to the courts and society regarding understanding and predicting human behavior, but the legal arena is different from the traditional areas in which psychology has been practiced. It is very important for any psychologist wishing to practice forensic psychology to know the applicable guidelines, standards, laws, and regulations regarding forensic psychology, and to understand the legal processes. The hope and expectation is this chapter may serve to help education psychologists wishing to enter the forensic arena and may serve to facilitate the future use of psychological knowledge by the courts.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (3rd ed). Washington, DC: Authors.
- American Psychological Association. (2007). Record keeping guidelines. *American Psychologist*, 62, 993–1004.
- American Psychological Association (2010a) Ethical principles of psychologists and code of conduct Retrieved from <http://www.APA.work/ethics/code/index.aspx> (2002, Amended June 1, 2010).
- American Psychological Association. (2010b). Guidelines for child custody evaluations in family law proceedings. *American Psychologist*, 65, 863–867.
- American Psychological Association. (2013a). Specialty guidelines for forensic psychology. *American Psychologist*, 68, 7–19.
- American Psychological Association. (2013b). Guidelines for psychological evaluation in child protection matters. *American Psychologist*, 68, 20–31.
- American Psychological Association. (2013c). Guidelines for assessment of and intervention with persons with disabilities. *American Psychologist*, 67, 43–62.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders, fifth edition (DSM-5)*. Arlington, Virginia: 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(1), 84–94.
- Blau, T. H. (1998). *The psychologist as an expert witness* (2nd ed). New York: Wiley.
- Block, R. & Berman, P. C. (2015) Forensic psychology: Practice, informed consent and record keeping. *Maryland Psychological Association (MPA) Workshop, June 19, 2015, Columbia, Maryland*.
- Brodsky, S. L. (2013). *Testifying in court in the: Guidelines and maxims for the expert witness* (2nd ed). Washington, DC: American Psychological Association.
- Brodsky, S. L., & Dattilio, T. G. (2016). *The expert expert witness: more maxims and guidelines for testifying in court* (2nd ed). Washington, DC: American Psychological Association.

- Bush, S. S., & Martin, T. A. (2010). Privacy, confidentiality, and privilege in forensic neuropsychology. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (pp. 235–244). New York: Springer.
- Bush, S., Connell, M. A., & Denney, R. L. (2006). *Ethical practice in forensic psychology: A systematic model for decision making*. Washington, DC: American Psychological Association.
- Bush, S., Ruff, R., Troster, A., Barth, J., Koffler, S., Pliskin, N., Reynolds, C., & Silver, C. (2005). Symptom validity assessment: Practice issues and medical necessity-NAN Policy and Planning Committee. *Archives of Clinical Neuropsychology*, 20(4), 419–426.
- Bush, S. S., Rapp, D. L., & Ferber, P. S. (2010). Maximizing test security in forensic neuropsychology. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (2nd ed, pp. 177–196). New York: Springer.
- Crown, B. M., Fingerhut, H. S., & Lowenthal, S. J. (2010). Conflicts of interest and other nettlesome pitfalls for the expert witness. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (2nd ed, pp. 245–275). New York: Springer.
- Denney, R. L., & Tyner, E. A. (2010). Criminal law, competency, insanity and dangerousness: competency to proceed. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (pp. 211–234). New York: Springer.
- Foote, W. E. (2006). Ten rules: How to get along better with lawyers and the legal system. In S. N. Sparta, & G. Koocher (Eds.), *Forensic Mental Health Assessment of Children and Adults* (pp. 64–73). New York: Oxford University press.
- Franzen, M. (2010). Neuropsychological evaluations in the context of civil competency decisions. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (pp. 197–210). New York: Springer.
- Freud, S. (1906). Psychoanalysis and the certainty of truth in court of law. *Clinical papers and papers on technique, Collected papers* (1959) 2, 13–24. New York: Basic Books.
- Frye v. United States, 293 F. 1013 (DC Civ 1923).
- Gottfried, E. D., Schenk, A. M., & Vitacco, M. J. (2016). Retrospectively assessing for feigning in criminal responsibility evaluations: Recommendations for clinical practice. *Journal of Forensic Psychology Practice.*, 16(2), 118–128.
- Greenberg, S. A., & Shulman, D. W. (1997). Irreconcilable conflict between therapeutic and forensic roles. *Professional Psychology: Research and Practice*, 28(1), 50–57.
- Greenberg, S. A., & Shulman, D. W. (2007). When worlds collide: Therapeutic and forensic roles. *Professional Psychology: Research and Practice*, 38(2), 129–132.
- Grisso, T. (2010). Guidance for improving forensic reports: A review of errors. *Journal of Forensic Psychology*, 2, 102–115.
- Guthiel, T. G. (2009). *The psychiatrist as expert witness*. (2nd ed). Arlington, VA: American Psychiatric Association.
- Guthiel, T. G., & Dattilio, F. M. (2007). *Practical approaches to forensic mental health testimony*. Baltimore, MD: Lippincott Williams & Wilkins.
- Hartlage, L. C. (2003). Neuropsychology in the courtroom. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (pp. 315–333). New York: Springer.
- Hartlage, L. C., & Stern, B. H. (2010). Neuropsychology in the courtroom. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (pp. 29–308). New York: Springer.
- Hartlage, L. C., & Williams, B. L. (2010). Depositions. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (pp. 277–290). New York: Springer.

- Heilbrun, K., Marczyk, G., DeMatteo, D., Zillmer, E., Harris, J., & Jennings, T. (2003). Principles of forensic mental health assessment: Implications for the neuropsychological assessment in the forensic context. *Assessment, 10*, 329–343.
- Hopwood, C. J., & Sellbom, S. (2013). Implications of DSM-5 personality traits for forensic psychology. *Psychological Injury and Law, 6*(4), 314–323.
- Horton, A. M., Jr., & Hartlage, L. C. (Eds.). (2003). *Handbook of forensic neuropsychology*. New York: Springer.
- Iverson, G. L. (2003). Detecting malingering in civil forensic evaluations. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology*. (pp. 137–177). New York: Springer.
- Iverson, G. L. (2010). Detecting exaggeration, poor effort and malingering in neuropsychology. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (2nd ed, pp. 91–136). New York: Springer.
- Kurke, M. I. (1980). Forensic psychology: A threat and a response. *Professional Psychology: Research and Practice, 11*, 72–77.
- Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research and Practice, 34*95, 491–498.
- Larrabee, G. L. (2005). Assessment of malingering. In G. L. Larrabee (Ed.), *Forensic neuropsychology: A scientific approach* (pp. 115–158). New York: Oxford.
- Malek-Ahmadi, M., Erickson, T., Puente, A. E., Pliskin., & Rock, R. (2012). The use of psychometrists in clinical neuropsychology: History, current status and future directions. *Applied Neuropsychology: Adult, 19*, 26–31.
- Malone, D. M., & Zweier, P. J. (2014). *Effective expert testimony* (3rd ed). Notre Dame, IN: National Institute for Trial Advocacy.
- Melton, G. B., Petrila, J., Poythress, N., & Slobogin, C. (2007). *Psychological evaluation for the courts: A handbook for mental health professionals and lawyers* (3rd ed). New York, NY: Guilford Press.
- Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*, 1094–1102.
- Munsterberg, H. (1908). *On the witness stand*. New York: Doubleday.
- Otto, R. K., & Heilbrun, K. (2002). The practice of forensic psychology: A look towards the future in light of the past. *American Psychologist, 57*(1), 5–18.
- Otto, R. K., DeMint, R., & Boccaccini, M. T. (2014). *Forensic reports and testimony: A guide for effective communication for psychologists and psychiatrists*. New York, NY: Wiley.
- Pirelli, G., Otto, R. K., & Estoup. (2016). A. Using the internet and social media data as collateral sources of information in forensic evaluations. *Professional Psychology: Research and Practice, 47*(1), 12–17.
- Project on Scientific Knowledge and Public Policy (June 2003). *Daubert: The most influential Supreme Court ruling you've never heard of*. Retrieved from <http://www.defendingscience.org/courts/Daubert-report-excerpt.cfm> (15.01.07).
- Reynolds, C. R. (1998). *Detection of malingering during head injury litigation*. New York, NY: Plenum Press.
- Reynolds, C. R., & Horton, A. M., Jr. (2012). Clinical acumen, common sense, and data related decision-making in the assessment of dissimulation during head injury litigation. In C. R. Reynolds, & A. M. Horton, Jr. (Eds.), *Detection of malingering during head injury litigation* (2nd ed, pp. 351–370). New York, NY: Springer.
- Reynolds, C. R., & Livingston. (2012). *Mastering modern psychological testing: Theory & methods*. Upper Saddle River, New Jersey: Pearson Education, Inc.

- Rogers, R. (Ed.), (2008). *Clinical assessment of malingering and deception* (3rd ed). New York, NY: Guilford Press.
- Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of pediatric neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York, NY: Oxford.
- Watson, J. B. (1913). Psychology as the behaviorist sees it. *Psychological Review*, 20, 159–177.
- Weithorn, L. A. (1984). Children's capacities in legal contexts. In N. D. Reppucci, L. A. Weithorn, E. P. Mulvey, & J. Monahan (Eds.), *Children, mental health, and the law* (pp. 25–55). Beverly Hills, CA: Sage.
- Weithorn, L. A. (2006). The legal contexts of forensic assessments of children and families. In S. N. Sparta, & G. P. Koocher (Eds.), *Forensic mental health assessment of children and adolescents* (pp. 11–29). New York: Oxford.
- World Health Organization. (2005). International classification of diseases-10th revision-clinical modification (*ICD-10-CM*). Geneva: Author.
- Weiss, K. J., & Watson, C. (2015). *Psychiatric expert testimony: Emerging applications*. New York, NY: Oxford University Press.
- Zillmer, E. A., & Green, H. K. (2006). Neuropsychological assessment in the forensic setting. In R. P. Archer (Ed.), *Forensic uses of clinical assessment Instruments*. (pp. 209–227). Mahwah, NJ: Lawrence Erlbaum Associates.

## Further reading

- American Bar Association & American Psychological Association. (2008). *Assessment of older adults with diminished capacity: A handbook for psychologists*. Washington, DC: American Bar Association and American Psychological Association.
- American Psychological Association. (2003). Guidelines on multiple multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist*, 58, 377–402.
- American Psychological Association. (2004). Guidelines for psychological practice with older adults. *American Psychologist*, 59(4), 236–260.
- American Psychological Association. (2012a). Guidelines for evaluation of dementia and age-related cognitive change. *American Psychologist*, 67, 1–9.
- American Psychological Association. (2012b). Guidelines for practice of parent coordination. *American Psychologist*, 67, 63–71.
- American Psychological Association. (2012c). Guidelines for psychological practice with lesbian, gay and bisexual clients. *American Psychologist*, 67, 10–42.
- Daubert v. Merrell Dow. (1993). 509 US 579.
- Horton, A. M., Jr., & Hartlage, L. C. (Eds.), (2010). *Handbook of forensic neuropsychology* (2nd ed). New York: Springer.

# Fairness in psychological testing

19

Zarui A. Melikyan<sup>1</sup>, Anna V. Agranovich<sup>2</sup> and Antonio E. Puente<sup>3</sup>

<sup>1</sup>University of California Irvine, Institute for Memory Impairments and Neurological

Disorders, Irvine, CA, United States, <sup>2</sup>Department of Physical Medicine and

Rehabilitation, Johns Hopkins University School of Medicine, Baltimore, MD, United

States, <sup>3</sup>University of North Carolina Wilmington, Department of Psychology, Wilmington,

NC, United States

## Defining fairness in psychological testing

For many decades, psychological testing research suggested that cognitive, intelligence, and personality tests yield quite different results when administered to middle-class individuals of European decent (often referred to as “majority group”) versus diverse or minority groups (Boone, Victor, Wen, Razani, & Pontón, 2007). Disparities in test performance were found among ethnic and racial groups living in the same country (Byrd, Touradji, Tang, & Manly, 2004; Manly, Byrd, Touradji, & Stern, 2004), individuals of the same ethnic background coming from different countries and now living in the U.S. (Byrd et al., 2004; Puente, Perez-Garcia, Vilar-Lopez, Hidalgo-Ruzzante, & Fasfous, 2013), and individuals of the same ethnicity and language living in different countries (Buré-Reyes et al., 2013). This includes English-speaking countries such as United States, New Zealand, and Australia (Barker-Collo, 2001; Cruice, Worrall, & Hickson, 2000). Demographic (Mungas, Reed, Farias, & Decarli, 2009), health (Noble, Manly, Schupf, Tang, & Luchsinger, 2012), linguistic, sociocultural and socioeconomic variables (Boone et al., 2007; Manly et al., 2004) explain a significant part of variance in test performance across groups. Furthermore, sociocultural factors account for greater variance in test performance than neurological conditions (Rivera Mindt, Byrd, et al., 2008; Saez et al., 2014). If unaccounted for, these sources of variability introduce bias in testing (Robertson, Liner, & Heaton, 2009; Rosselli & Ardila, 2003). For test-takers from cultural minority groups, testing bias may result in inappropriate diagnosis, treatment, placement, or denial of services/positions (Dilworth-Anderson et al., 2008).

Bias in psychological testing may be introduced by a number of characteristics, including cultural relevance of a construct being tested, meanings of test items in different cultures and values of particular responses, communication conventions, and modes of knowing such as individual versus collective, or process versus object

([Ardila, 2005](#)). Majority of tests have been developed and standardized for middle-class Caucasian, monolingual, English-speaking individuals from North America. Therefore it is not surprising that members of the majority culture get higher scores while the aptness of tests for minorities becomes questionable ([Harris, Tulsky, & Schultheis, 2003](#)). In addition, psychologists often belong to and interpret test results from the perspective of “majority” culture without taking into account cultural specifics of a test-taker.

Therefore fairness in testing implies taking into account group and individual cultural differences at each stage of the evaluation to minimize bias and offer valid test results and conclusions. Fairness in psychological testing means providing equal opportunity for all the test-takers to demonstrate their true performance on the measured construct. Fairness offers comparable validity across groups within the population of interest. Fair testing strives to diminish or avoid factors that may produce variance in test performance that is unrelated to the measured construct (i.e., bias). These factors include:

1. *Test content* that may evoke different performance in examinees from different groups due to irrelevance of testing materials and procedures. Test content may also contribute to differential motivational involvement and emotional response by culturally diverse test-takers.
2. *Context of testing*, which include stereotype threat ([Steele & Aronson, 1995](#)), interpersonal relationships between test administrator and test-taker, appropriateness of language used in the examination, clarity of instructions, and complexity of language demands.
3. *Responses to tests*, which may vary due to language proficiency, disability, and cultural or individual characteristics of a test-taker, which may evoke an unintended response.
4. *Familiarity* with the concepts/constructs being tested, which may be limited for recent immigrants, disadvantaged populations, and individuals with limited access to quality education.

According to the [APA Standards for Educational and Psychological Testing \(2014\)](#), fairness in testing is provided by: (1) fair treatment of a test-taker during testing, (2) avoiding measurement bias, (3) providing access to the constructs as measured, and (4) practicing individualized approach to the interpretation of test performance.

*Fair treatment* during testing presumes adherence to standardized testing and scoring procedures for all the test-takers. However, flexibility in the evaluation may be needed when standard procedures will prevent some test-takers from accurately demonstrating their ability. This may be applicable for individuals with limitations in function, the socio-economically disadvantaged, or for not fully acculturated individuals. Psychologists need to ensure equal level of familiarity of test-takers with any technical tools that may be used in the assessment and need to be well qualified to ensure correct evaluation procedures and adequate communication throughout the testing.

*Measurement bias* occurs when equally able test-takers differ in their probability to correctly perform on the test because of their group membership. A related issue is the extent to which the construct of the test has the same meaning for representatives of different groups.

*Access to measured construct* allows all the test-takers to show their abilities on the measured construct without being disadvantaged by characteristics that are irrelevant to the measured construct but are necessary in order to take the test, such as visual or hearing acuity or language proficiency.

*Individualized approach to interpretation of test scores* requires consideration of the multitude of unique characteristics of each test-taker rather than the test-taker's group membership.

The term “minority” refers to a category of people different in certain respects from the majority group that is in a position of social power in a society. Minority groups often constitute vulnerable populations that warrant a different approach to testing. For the purposes of psychological assessment, *minority* is defined as people diverse from the majority with respect to cultural, linguistic, ethnic, racial, educational, socioeconomic background, age, and/or disability status. One of the evaluation goals should be to ensure that these variables do not lead to measurement error with possibly significant social and individual consequences. Since cultural, linguistic, racial, and ethnic diversity groups predominate among minorities in terms of number and need for services, the majority of discussion in this chapter is dedicated to these groups.

## Importance of fairness in testing minority individuals

Minority population is predicted to significantly increase among psychology clientele in the near future, given the growth of this population due to migration and natural increase. According to [United Nations Department of Economic and Social Affairs Population Division \(2016\)](#), in 2015 3.3% of the world's population lived outside of their country of origin. In 2015 minorities constituted 38.7% of the U.S. population and the number is projected to grow so that by 2044 the United States would become a “majority–minority country” ([Colby & Ortman, 2014](#)). Over recent decades, there has been a trend toward growing awareness of available resources, trust in mental health professionals, and accessibility of healthcare through social programs, including the Affordable Care Act ([Krahn, Walker, & Correa-De-Araujo, 2015](#)) in minorities with increased risk for physical and mental disorders.

Minority groups are internally diverse ([Llorente, 2008](#)). Although minorities are usually compared to the white North-American population, differences among the minority groups on test performances are quite significant in verbal and nonverbal tasks ([Rosselli & Ardila, 2003](#)), tests of attention ([Byrd et al., 2004](#)), memory ([Ostrosky-Solis & Ramirez, 2004](#)), visual–perceptive ability ([Ostrosky-Solis & Ramirez, 2004](#)), cognitive flexibility ([Razani, Burciaga, Madore, & Wong, 2007](#)), abstract reasoning and category fluency ([Manly, Jacobs, Touradji, Small, & Stern, 2002](#); [Touradji, Manly, Jacobs, & Stern, 2001](#)), inhibition ([Razani et al., 2007](#)), intelligence ([Razani, Murcia, Tabares, & Wong, 2006](#)), and timed test performance ([Agranovich & Puente, 2007](#)).

Nation, race, ethnicity, and language are commonly used as proxies of culture, defined as “the embodiment of a worldview through learned and transmitted beliefs, values, and practices, including religious and spiritual traditions. It also encompasses a way of living informed by the historical, economic, ecological, and political forces on a group” (American Psychological Association, 2003). Although contemporary men are becoming more and more culturally homogeneous due to diffusion of characteristics across cultures by means of education and communication technology, culture still dictates what is relevant and significant, and leads to development of specific behaviors and abilities (Ardila, 2005; Glymour & Manly, 2008). Culture shapes mental and cognitive processes (Fasfous, Hidalgo-Ruzzante, Vilar-López, Catena-Martínez, & Pérez-García, 2013; Sternberg & Grigorenko, 2004), personality (Church, 2016; McCrae & Terracciano, 2005), brain morphology and activation patterns (Paige, Ksander, Johndro, & Gutchess, 2017; Park & Huang, 2010), as well as allele variations (Chiao & Blizinsky, 2010; Way & Lieberman, 2010), even when accounting for education and relevant demographics.

Approach to psychological testing may vary, depending on whether cognition is viewed as a product of interaction between biological and sociocultural factors, or as a fundamentally universal function, independent of the cultural and environmental context in which it develops (Luria, 1976; Nell, 2000; Vygotsky, 1978). North-American psychological tradition was mostly influenced by the latter view, which did not consider cultural, linguistic, educational and socioeconomic factors in test construction or their application for clinical and research purposes (Ardila, Rosselli, & Puente, 1994; Sifers, Puddy, Warren, & Roberts, 2002). Although this testing approach has been presented as suitable for representatives of any culture and background (Ardila, 1996), it is not surprising that tests designed for urban, middle-class, English-speaking, educated individuals yield different results in minority populations.

The importance of cultural considerations for evaluation of diverse individuals has been recognized by professional organizations. The American Psychiatric Association (APA) included the “Cultural Formulation” chapter in the *DSM-IV* (2000) and *DSM-5* (2013). The APA developed “Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations” (2003) and “Multicultural Guidelines: An Ecological Approach to Context, Identity, and Intersectionality” (2017b). The APA’s “Ethical Principles of Psychologists and Code of Conduct” (2017a) state that use of inappropriate measures for culturally different individuals and/or refusing diversity population services that are beneficial for them is unethical. American Academy of Clinical Neuropsychology (AACN) created a *Relevance 2050 Initiative* in anticipation that by 2050, 60% of Americans would be “untestable” with existing monolingual and monocultural instruments, which would make neuropsychology irrelevant in the healthcare market and would create a social justice issue.

A culturally fair approach is essential for valid evaluation of minorities. Such approach identifies and evaluates both universal and unique psychological characteristics (Perez-Arce, 1999). Universal characteristics and processes are largely determined by genetic factors (Kennepohl, 1999), whereas unique characteristics

develop during the process of social interaction in a particular cultural, social, and historical context (Vygotsky, 1978). Put another way, cognitive processes are similar between cultures, but their manifestations in each culture are different (Sternberg & Grigorenko, 2004). More research is needed to elucidate mechanisms of these differences and possibilities for more fair assessment. This chapter offers a review of variables contributing to differences in test performances among cultural groups, approaches to account for them, and suggestions for future directions.

## Variables contributing to test performance in minorities

### ***Psychological construct being tested and test content***

Currently, there is shortage of tests appropriate for use with minorities. Hence, psychologists are forced to use tests with limited cultural relevance (Zebrowski, Vega, & Llorente, 2015) and questionable validity of test results (Daugherty, Puente, Fasfous, Hidalgo-Ruzzante, & Pérez-Garcia, 2017; Norman et al., 2011). For culturally relevant evaluation of the desired construct, the following options for test selection should be considered: (1) most valid test for the minority group the test-taker belongs to, with respect to psychological construct(s) tested, test procedures, and stimulus materials; (2) culturally competent translation and adaptation of already existing test; (3) test developed for one or several related minority groups.

1. It is assumed that psychological tests are measuring constructs that are relevant and important in order to succeed in society. However, the validity of those constructs for diverse groups was rarely considered, and *de facto* tests tend to measure concepts relevant for the culture in which they were developed, with questionable relevance of those concepts to dissimilar cultures. For example, in Western cultures, speed and quality of performance are equally important, yet in some Eastern and Southern cultures quality of performance overrides speed. Therefore, timed tests may not be “culture-fair” in the cultures where efficiency or time pressure are irrelevant constructs (Agranovich, Panter, Puente, & Touradji, 2011). At the same time, individuals from non-Western cultures outperform Westerners on the tests that are relevant to their culture and daily practices. For example, perceptual constancy (stability of perception regardless of changes of actual characteristics of stimuli) is more accurate in non-Western low-schooled societies than in the literate and westernized individuals (Ardila, 1996).

Northern-American standardized testing procedures may not always correspond to typical experiences in other cultural groups. Alternative approaches such as process approach, Luria's (1980) hypothesis testing, or creative testing based on familiarity with cultural schemas, may be more relevant for assessment of minorities.

When having to use the existing tests (e.g., for test-takers culturally similar to test-maker), it is important to consider the validity of construct(s), stimulus materials and test procedures for the test-taker's group (Ardila, 2007a).

2. Test translation and adaptation are the most common practices used for cross-cultural comparisons, accommodation of language and cultural needs of test-takers, and time and cost reduction in test development. However, these practices are based on assumptions that do not always hold true: (1) existing tests are optimal for assessment of the domain

regardless of culture; (2) translated tests are equivalent to those in the original language; and (3) psychometric properties of the original test transfer into the new language (Ardila, Rodriguez-Menendez, & Rosselli, 2002).

Test items of similar content are not necessarily equally culturally relevant or assess cognitive processes valued in different cultures, therefore test adaptations are performed to increase relevance and preserve conceptual equivalence. Adaptation takes the form of *accommodation* when the test construct is retained, and introduced changes allow for performance comparability (e.g., magnification of stimulus materials to compensate for decrease in visual acuity). *Modification* is a form of adaptation when a measured construct is transformed according to cultural and linguistic demands to get a measure of a somewhat different but appropriate test construct (e.g., switching between colors while following numerical sequence in more culturally fair Color Trails Test (D'Elia, Satz, Uchiyama, & White, 1996) instead of following alpha-numerical sequence in the original Trail Making Test B). Nonverbal stimuli often also need cultural adaptation and establishing item equivalence (Rosselli & Ardila, 2003). Differences across cultures have been found in performance on tests of depth perception, design copy, perception of overlapping figures, tactual or rhythm tests (Ardila & Moreno, 2001). Item equivalence may be established using judgment of culturally and linguistically competent professionals or via statistical procedures comparing the test results across groups.

Translation and adaptation should adhere to existing guidelines (International Test Commission, 2016) for translation, item format and appearance, stimulus materials, cultural relevance and specificity (Hambleton & Zenisky, 2011), and should be provided by professionals experienced in psychological, linguistic, and cultural issues. It can be accomplished through translation/back translation and altering or removing culturally relevant elements of the test (van de Vijver & Leung, 1997). Site translation during testing by a bilingual professional or having an untrained individual conduct the evaluation simply because they know the language of the test-taker are discouraged (Candelaria & Llorente, 2009). Using the test in a new linguistic culture requires establishing its reliability and validity, and assuring equivalence of original and translated tests and their scores (American Educational Research Association 2014; Candelaria & Llorente, 2009).

3. More rigorous and more resource-consuming approach would be to develop tests that account for cultural variables. Tests that take into account unique group characteristics, are best for assessment within a group or for similar groups, but not appropriate for between-group comparisons (Ardila, 2005). Psychometrically matched test forms for different cultures and languages, such as Spanish and English Neuropsychological Assessment Scales (SENAS), allow for valid cross-cultural comparisons (Marin, 1992). The so-called “culture-fair” or “culture-free” tests intend to assess constructs common to diverse groups using test materials and procedures relevant across groups and cultures and therefore are appropriate for cross-cultural comparisons. Though it is not possible to completely eliminate cultural influence in the tests, a list of relatively culture-fair tests include: WHO/UCLA Test Battery (Maj et al., 1993), Cross-Linguistic Naming Test (CLNT; Ardila, 2007b), European domain-specific computerized battery for cross-cultural comparisons (EMBRACED; Ibanez-Casas, Daugherty, Leonard, Perez-Garcia, & Puente, 2017), Tower of London—Drexel University (ToLDx); (Culbertson & Zilmer, 2001). Yet, further validation showed cultural differences in some of those tests, for example, Children's and Adult Color Trails Test (Agranovich et al., 2011; Fasfous, Puente, et al., 2013).

In situations when formal cognitive assessment might be culturally biased, or when assessing individuals with limited education and/or low socioeconomic status,

self- and informant-report may prove to be more ecologically valid than standardized tests (Ardila, 2005).

### **Test norms**

The normative approach to test results evaluation entails comparison of a test-taker's performance against that of healthy peers comparable to the test-taker on aspects crucial for test performance (Mitrushina, Boone, Razani, & D'Elia, 2005). Most of the existing norms were obtained from convenience samples: urban, middle-class, educated, English-speaking individuals of European descent, with minorities not being fully represented. These norms account for characteristics relevant for test performance in this select group: age, education, and sometimes gender or ethnicity (i.e., Caucasians and African Americans, Mitrushina et al., 2005). Other variables, most relevant for minority test performance, such as socioeconomic status (SES), literacy level, rural versus urban dwelling, culture, or language have not been systematically studied (Perez-Arce, 1999). Accurate assessment of minorities requires norms appropriate for the tested population (Lucas et al., 2005) with relevant parameters (e.g., education brackets, SES) taken into account.

Existing norms should be used carefully for minority evaluations because they may not reflect demographic and cultural parameters critical for test performance in this group. This may lead to inaccurate test data interpretation (Harris & Llorente, 2005). Even minority norms may not be completely representative of test-taker. For example, standardization group from the same country as the test-taker may still differ from him/her with respect to demographic and cultural variables (Llorente, 2008). Racial/ethnic norms may ignore variables relevant for test performance in minorities, such as SES, education, acculturation (Manly, 2006), for which race/ethnicity serves as proxy. It is most informative when one of the important variables—education—is described by its quality, not quantity, as it is usually done in the norms. Racial/ethnic norms ignore within racial/ethnic group variability that is greater than between group variability. In addition, race/ethnicity specific norms may set lower cut-offs, which may lead to ignoring existing deficits and underserving populations in need (Manly & Echemendia, 2007). In this situation, Ardila (2007a) suggests using the norms closely matching the characteristics of test-taker and be aware of the sources of variation in test performance in different cultural groups.

Norms specifically developed for minority groups are currently insufficient, but represent the most rigorous approach. Due to the multitude of cultures and other parameters influencing test performance, it is impossible or impractical to obtain norms for all the existing cultural groups. Yet, it may be reasonable to attempt to develop norms for clusters of groups that are culturally similar.

Diagnostic questions may help deciding which norms to use in a particular case. If the goal of an evaluation is to compare a test-taker to demographically matched majority population, then use of the standard North American norms may be appropriate. However, if the evaluation is done for diagnostic purposes to identify a particular impairment, then norms from a group closely matched to the particular

test-taker should be used (Manly & Echemendia, 2007). Manly and Echemendia (2007) suggests using norms that are of most benefit and least cost to the test-taker. For example, given the benefits of early detection and intervention of TBI-associated cognitive impairment, one would want to have increased diagnostic sensitivity; whereas in the context of a criminal forensic case, norms that provide greater specificity would be more appropriate.

### ***Acculturation and assimilation***

Acculturation and assimilation are processes of transmission of cultural phenomena between minority and majority populations, which reflect immersion of an individual in a culture and the degree of its internalization. *Assimilation* is the cultural absorption of the minority group by the majority, in which the assimilated group loses its cultural characteristics. *Acculturation* is transmission of cultural features between groups, in which each of them adopts some of the features of the other group but still remains distinct. Measurement of acculturation to both the majority culture as well as a culture of origin is warranted (Lopez-Class, Gonzalez Castro, & Ramirez, 2011). Acculturation can occur in various degrees along a continuum: from minimal in those who live in ethnic neighborhoods and speak native language, to complete immersion in the majority culture. It involves learning the language, history and traditions of new culture, changing one's own behaviors, norms, values, worldview, and interaction patterns (Marin, 1992).

Culture dictates attitudes and behaviors during testing, including test-wiseness, proficiency in explicit and implicit requirements for test performance, motivation, feelings of insecurity, perception of possible discrimination, or frustration with time- and effort-consuming evaluation (Perez-Arce, 1999). Individuals from Western cultures usually appreciate that testing is a challenge and that fast and best performances are crucial (Ardila, 2005; Manly, 2006). For Hispanics, it may be more important to establish a relationship with a tester, and Russians tend to value quality of performance more highly than efficiency (Agranovich et al., 2011).

Greater acculturation is generally associated with higher performance on tests of global functioning, executive function, naming, verbal fluency, learning and memory, and processing speed (Arentoft et al., 2012; Coffey, Marmol, Schock, & Adams, 2005; Manly et al., 2004), although results that report lack of differences have also been published (Boone et al., 2007). Personality testing may also be affected by acculturation to the extent that the differences may lead to different diagnoses and management decisions (Cueilar, 2000).

Test-taker's level of acculturation should be evaluated prior to testing to tailor the evaluation or refer to another specialist who is more proficient in the test-taker's culture or language. In nonimmigrant groups, acculturation is best assessed by segregation level of schooling and current and/or childhood residential segregation (Manly, 2006). Among immigrant groups, acculturation is best assessed by years in the country, timing of immigration, English language proficiency, bilingualism/most frequently used language at work/home, level, quality, and country of education and employment, and social contacts. Such information may be collected from

a test-taker and/or informant (Llorente, 2008; Zebrowski et al., 2015) as well as by formal evaluation of language proficiency and acculturation (Marin & Gamba, 1996; Stephenson, 2005; Unger et al., 2002).

Tests selection should be appropriate for the culture of the test-taker. Depending on the level of acculturation, most appropriate norms should be either derived from immigrant minorities or from those who were born in the country (Llorente, Taussig, Satz, & Perez, 2000). Test results should be interpreted with respect to individual's test-wiseness, motivation, and other variables related to the level of acculturation. To increase the ecological validity of recommendations, test findings should be placed in social and cultural context of the individual and his/her degree of acculturation (Dana, 1993).

### ***Communication and language***

Effective verbal and nonverbal communication is instrumental for a better insight into the test-taker's emotional and motivational state and, ultimately, valid assessment. Nonverbal communication differs quite a bit across cultures, with some cultures relying more heavily on nonverbal communication than others. Meaning of nonverbal communication signs may vary across cultures and may even carry opposite meaning (e.g., head nodding signifies agreement in most European countries, but has direct opposite meaning in Bulgaria).

Tests may elicit some unintended responses, or responses typical in some groups may systematically alter the results: for example, providing expected answers appropriate for majority culture instead of those characteristic of the test-taker. In certain cultures, verbosity may be considered rude; in others, it may be an indicator of mental abilities and friendliness. At times, the test format may require a level of language proficiency that can be challenging for a test-taker.

Language background accounts for variance in test performances between different ethnicities (Latino/a and non-Hispanic White) and between bilingual and monolingual groups of the same ethnicity (Gasquoine, Croyle, Cavazos-Gonzalez, & Sandoval, 2007; Mungas, Reed, Haan, & González, 2005) even after accounting for relevant demographic variables (Harris & Llorente, 2005). Proposed reasons for different performance of bilinguals and monolinguals involve two competing mechanisms: interference between two languages and reduced frequency of language-specific use (Rivera Mindt, Arentoft, et al., 2008).

To preserve testing validity in individuals for whom English is not the first or native language, it is imperative to determine the language in which testing should be done. There is no clear evidence whether testing in a test-taker's native versus acquired language yields similar or different results, and if differences exist, what they might be. For bilinguals, linguistic proficiency in both languages should be assessed. Bilingual individuals speak different languages in different contexts and with different people; fluency and competency in each language may vary and they may perform better if tested in one language over the other. Bilingualism can be viewed as a spectrum: on one extreme are people who are proficient in one of the two languages and have basic knowledge of the other, and on the other extreme

are bilinguals with equal proficiency in both languages (Albert & Obler, 1978). APA's Guidelines (2003) state that test-takers should be tested in appropriate language and, if that is not possible, be referred to a provider who speaks the language. If that is not done, testing may lead to invalid scores and misclassification of a minority individual as impaired. However, guidelines and instruments to evaluate language proficiency and bilingualism are currently insufficient (Elbulok-Charape, Rabin, Spadaccini, & Barr, 2014).

Language proficiency should never be assumed based on the observed social communication, because language processing in social situations with ample non-verbal situational cues is quite different from the testing situation, where such cues are limited. Because language proficiency may vary depending on a skill assessed, evaluation of reading, writing, speaking and comprehension is preferred. Accurate assessment includes both subjective and objective measures. The former include questions about language proficiency and usage, literacy, education and academic achievements in a particular language, along with measures of acculturation (Llorente, 2007). Objective measures consist of verbal fluency and naming tests (e.g., Boston Naming Test; Rosselli et al., 2002). Multidomain evaluation (e.g., Woodcock Munoz Language Survey-Revised) does not provide any advantage in language proficiency evaluation compared to individual tests (Miranda et al., 2016).

Language dominance index can be calculated based on self-report and objective measures. English-dominant bilinguals should be tested in English by an English-speaking psychologist, while a bilingual examiner will help establish the degree of bilingualism. Non-English-dominant bilinguals should be evaluated in their dominant language by a psychologist fluent in that language and cognizant of the appropriate ethnic and cultural background. For a relatively balanced bilingual, an acculturation measure can help determine the language of examination (Ponton, 2001). However, testing in both languages is preferred in order to capture information that is more readily available in one language than the other (Paradis, 2008). Due to the possibility of language interference, it is advisable that testing in different languages be done in separate blocks. Interpretation of results should consider disadvantage of bilinguals compared to monolinguals in producing low frequency words and/or on tasks that increase language interference (Ivanova & Costa, 2008). Some language-based tasks (e.g., verbal memory) have not revealed effect of testing language and can be used with bilinguals with greater confidence (Gasquoine et al., 2007).

A separate task is to determine whether psychologist's language proficiency and familiarity with the culture are sufficient to test bilingual or non-English speaking test-taker (Ponton, 2001). Training and competency guidelines are not sufficient, and it is left to a psychologist's ethical and professional judgment to decide whether he/she is competent to conduct the examination in a particular language. The literature mostly suggests self-examination: whether the psychologist is a native speaker or a balanced bilingual; whether his/her language proficiency is equivalent to those who completed graduate studies at a foreign university or lived in the country for several years (Rivera Mindt, Arentoft, et al., 2008). A psychologist who speaks the

same language as the test-taker but is from a different culture should be aware of linguistic regionalisms and cultural particularities. Alternatively, a team of primary examiner (qualified practitioner) and ancillary examiner (native in test-taker's language and culture and trained in principles of test administration) may be used (Woodcock & Munoz-Sandoval, 1996). Professionally trained native speakers would be ideal, but at least equal proficiency in language(s) and familiarity with culture(s) of the test-taker is required of a psychologist (Ardila et al., 2002). Taxonomy of linguistic competency is provided by American Council on the Teaching of Foreign Languages (ACTFL; American Council on the Teaching of Foreign Languages, 2012), National Standards for Foreign Language Learning in the 21st Century (NSFLL) or linguistic research (Savignon, 2007). Sociocultural competency guidance could be found in APA's Multicultural Guidelines on Education, Training, Practice, Research, and Organizational Change for Psychologists (2003).

Using interpreters for testing should generally be avoided (Candelaria & Llorente, 2009; Llorente, 2008). Changing the way the test is administered, by introducing an interpreter, requires re-standardization to assure validity and reliability of results (Melendez, 2001). Interpreters may alter questions and responses, distort important psychological meaning of the information, or may not be able to convey qualitative linguistic information, which is crucial, for example for patients with aphasia. Individuals with a preexisting relationship with a test-taker (e.g., friends, relatives, colleagues) should never be used to interpret, as their motivation to objectively transmit information, their level of language proficiency and acculturation, and potential lack of training in translation and ethics are likely to impact the quality of interpretation.

Evaluations using interpreters for forensic purposes are easily challenged in the court of law (LaCalle, 1987). Using interpreters is only acceptable when a test-taker speaks a rare language and extensive attempts to locate a psychologist who speaks that language have failed. In this case, high quality interpretative services with trained professional interpreters (preferably with a psychology or healthcare background) should be utilized (Paradis, 2008). The psychologist should educate the interpreter regarding the process of evaluation and the expectations for translation. Guidelines on how to work with interpreters are provided by Ponton and Corona-Lomonaco (2007) and the National Academy of Neuropsychology Position Paper (Judd et al., 2009). When interpreters are used, limitations to the validity of test results should be emphasized in the report.

Using norms for other languages or bilingual test norms, which are currently nonexistent, or "adjusting" test scores, are possible ways to account for testing in a foreign language and bilingualism (Ardila et al., 2002). The psychologist must make judgment on how well the tests and norms apply to the non-English speaking/bilingual examinee and make explicit comment in the report regarding potential limitations (Rivera Mindt, Arentoft, et al., 2008). Useful guidelines on how to prepare for and work with bilingual examinees are provided in several publications (Llorente, 2008; Paradis, 2008; Rivera Mindt, Byrd, Saez, & Manly, 2010).

## **Socioeconomic status and education**

SES is defined by the APA as social standing or class of an individual or group, and is often measured as a combination of education, income, and occupation. Inequalities in access to resources and issues related to privilege, power, and control are seen in groups with different SES ([American Psychological Association Task Force on Socioeconomic Status, 2007](#)). Traditional measures of SES, such as earnings and years of schooling, are often inappropriate for minorities ([Kaufman, Cooper, & McGee, 1997](#)). Instead, SES could be assessed through individual's assets, debt, use of public assistance, neighborhood-level indicators of income, ability to pay monthly bills, and home and/or vehicle ownership ([Diez Roux et al., 2001](#)). For the evaluation of minorities, it is essential to account for the effects of SES on test performance and to distinguish it from cultural impacts ([Betancourt & Lopez, 1993](#)).

Individuals with low SES cannot afford many opportunities that provide cognitively stimulating experiences and often demonstrate a low level of cognitive functioning throughout childhood, especially in executive functions and language ([Noble, McCandliss, & Farah, 2007](#)). Adult SES significantly correlates with verbal fluency, attention/concentration, memory, processing speed, executive functions; adjusting for SES has been shown to significantly attenuate group differences in cognitive performance ([Arentoft et al., 2015](#)).

One of SES' proxies, education, is an important contributor to test performance. Although some tests (e.g., language comprehension) are more sensitive to education than others (e.g., orientation), both verbal and nonverbal test scores correlate with education. The correlation is highest in low educated groups and goes down with increasing level of education. This effect may be explained by a relatively low test ceiling, differences in the abilities that tests assess (they may be relevant for educated individuals and irrelevant to illiterates), or by unfamiliarity with the concept being tested in illiterates ([Ardila, 1996](#); [Ostrosky-Solis, Ardila, & Rosselli, 1998](#)).

Education provides contents frequently included in cognitive tests, trains strategies of test-taking and information processing, and overall "test-wiseness"; it introduces attitudes toward and value of knowledge, testing, and test performance ([Gasquoine, 2009](#)). School can be considered a transnational culture in itself because it has the same fundamental aims and values regardless of location. Considering differences in learning opportunities and familiarity with the testing paradigm, it is not surprising that minorities score lower on the abilities related to schooling, while not showing differences in abilities to solve daily problems ([Ardila, 2007a](#)). Education is frequently confounded with cultural factors, hence the fact that what is frequently assumed to be a result of cultural differences may often be explained by educational disparities ([Byrd et al., 2004](#)).

Most commonly, the effect of education is accounted for by stratifying standardization samples by years of education. However, education-based norms may inflate impairment rates for minorities ([Ryan et al., 2005](#)) as they do not reflect the quality of education ([Cosentino, Manly, & Mungas, 2007](#)). Low quality education is associated with low cognitive performance and rapid decline ([Fyffe et al., 2011](#); [Manly, Schupf, Tang, & Stern, 2005](#)). Information on level of education achieved, school

location, student/teacher ratio, and rural versus urban school, should be used in evaluating quality of education (Manly, 2006).

Education quality can best be approximated by reading ability, as measured by the National Adult Reading Test (NART) or Reading subtests of Wide Range Achievement Tests, 4th Edition (WRAT-4; Fyffe et al., 2011). Reading level can be used in multiethnic and multilingual comparisons as it accounts for variation in test performance independently of race or language (Cosentino et al., 2007; Manly, 2006). Reading evaluation may be challenging for multilingual comparison due to linguistic differences. For example, a word-reading test in English requires a test-taker to read a series of phonologically regular and irregular words. In Spanish, however, there are no phonologically irregular words so entirely analogous test in unattainable. Instead, reading words with removed accentuation (e.g., Word Accentuation Test, WAT) would be appropriate (Del Ser, González-Montalvo, Martínez-Espinosa, Delgado-Villapalos, & Bermejo, 1997).

Another problem with using years of education is that they are not directly compatible across countries because education systems and years of education at each level may vary significantly. For example, high school education in Russia (and former Soviet Union) takes 10 or 11 years in comparison to 12 years in the U.S., but the curricula are comparable according to the World Evaluation Services. Until a couple of decades ago, there was no separate Bachelor's degree in Russia: most universities offered five-year long (for some majors, six) programs that were deemed equivalent to the North American Master's degree. Hence, familiarity with various educational systems would be required for equating levels of education. Norm development by level (rather than by years) of education may facilitate more accurate demographic corrections.

## Conclusions and future directions

With expected increase in the frequency of psychological services for minority individuals, it is essential that psychology prepares for this demographic shift and cultural challenge. Attention to cross-cultural issues in psychological assessment has historically been insufficient, and psychological tests developed in the U.S. and Europe have been assumed to be equally applicable across cultures. However, an increasing body of evidence shows differences in test performance across cultural groups. A fair approach to testing, which would allow everyone to demonstrate their true ability on the measured construct, is instrumental for the evaluation of individuals from cultural minority groups. This approach entails culturally relevant constructs, testing procedures, materials, and norms, as well as accounting for acculturation, language, education, and SES.

AACN's *Relevance 2050 Initiative* has identified directions for psychology to be able to offer culturally competent services, including research, development of new assessment methods, culturally sensitive training, mid-career supervision and clinical strategies available for professionals.

Future efforts to increase fairness in psychological assessment should include modifications to testing armamentarium and enhancement of cultural awareness among psychologists. Some of the existing tests can be translated, adopted, and validated according to strict guidelines. In other cases, new culturally appropriate tests would need to be developed. It is important to establish equivalency between different versions of a test, as well as between tests developed for different cultures, in order to measure similar constructs. Separate norms should be developed for groups of related cultures. Consideration of a test-taker's SES, education quality, level of acculturation, linguistic proficiency, and urban versus rural environment should be included in any case conceptualization.

When adopting testing materials or developing a new test, it is important to ensure that the measured construct is relevant in all the intended cultures and that a group of psychologists familiar with the culture and language confirms appropriateness of the test format and stimuli equivalence.

To continue increasing cultural awareness among psychologists, psychological study curriculums and continuing education should include courses on cross-cultural issues in assessment and intervention with minority populations. In-depth language courses, language proficiency exams, and internships abroad and/or with minority populations could help psychologists gain necessary cultural expertise. It is also essential to increase the number of diversity students in psychology training programs. Development and regular updates of specific guidelines for work with diverse populations would enhance psychological practice. Creating a network of providers who offer culturally sensitive services will be of great help for practitioners.

To promote the development of relevant knowledge, funding dedicated to research in this area should be available to scholars. Systematic studies of the impact of understudied cultural variables affecting test performances, such as SES, quality of education, urban and rural living environment, level of acculturation, and linguistic proficiency may also promote fair approach to testing. Cross-cultural validity of the psychological tests currently in use needs to be systematically studied and reported to help guide decision-making in neuropsychological assessment, from instrument selection to testing procedures and the interpretation of the results.

## References

- Agranovich, A. V., Panter, A. T., Puente, A. E., & Touradji, P. (2011). The culture of time in neuropsychological assessment: Exploring the effects of culture-specific time attitudes on timed test performance in Russian and American samples. *Journal of the International Neuropsychological Society*, 17(4), 692–701. Available from <https://doi.org/10.1017/S1355617711000592>.
- Agranovich, A. V., & Puente, A. E. (2007). Do Russian and American normal adults perform similarly on neuropsychological tests? Preliminary findings on the relationship between culture and test performance. *Archives of Clinical Neuropsychology*, 22(3), 273–282. Available from <https://doi.org/10.1016/j.acn.2007.01.003>.

- Albert, M., & Obler, L. (1978). *The bilingual brain*. New York: Academic Press.
- American Council on the Teaching of Foreign Languages. (2012). ACTFL proficiency guidelines 2012. Retrieved from [https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012\\_FINAL.pdf](https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf).
- American Educational Research Association American Psychological Association National Council on Measurement in Education & Joint Committee on the Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed). Washington, DC: Author.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed). Washington, DC: Author.
- American Psychological Association. (2003). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist*, 58(5), 377–402.
- American Psychological Association. (2017a). Ethical principles of psychologists and code of conduct (2003, amended June 1, 2010, and January 1, 2017). Retrieved from <http://www.apa.org/ethics/code/index.aspx>.
- American Psychological Association. (2017b). Multicultural guidelines: An ecological approach to context, identity, and intersectionality. Retrieved from <http://www.apa.org/about/policy/multicultural-guidelines.pdf>.
- American Psychological Association Public Interest Directorate. (2003). APA guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations. Retrieved from <https://doi.org/10.1037/e305082003-001>.
- American Psychological Association Task Force on Socioeconomic Status. (2007). *Report of the APA task force on socioeconomic status*. Washington, DC: American Psychological Association.
- Ardila, A. (1996). Towards a cross-cultural neuropsychology. *Journal of Social and Evolutionary Systems*, 19, 237–248.
- Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. *Neuropsychology Review*, 15(4), 185–195. Available from <https://doi.org/10.1007/s11065-005-9180-y>.
- Ardila, A. (2007a). The impact of culture on neuropsychological test performance. In B. P. Uzzell, M. O. Ponton, & A. Ardila (Eds.), *International handbook of cross-cultural neuropsychology* (pp. 23–45). Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.
- Ardila, A. (2007b). Toward the development of a cross-linguistic naming test. *Archives of Clinical Neuropsychology*, 22, 297–307.
- Ardila, A., & Moreno, S. (2001). Neuropsychological evaluation of Aruaco Indians: An exploratory study. *Journal of the International Neuropsychological Society*, 7, 510–515.
- Ardila, A., Rodriguez-Menendez, G., & Rosselli, M. (2002). Current issues in neuropsychological assessment with Hispanics/Latinos. In F. R. Ferraro (Ed.), *Minority and cross-cultural aspects of neuropsychological assessment* (pp. 161–179). Lisse, Netherlands: Swets & Zeitlinger.
- Ardila, A., Rosselli, M., & Puente, A. (1994). *Neuropsychological assessment of Spanish-speakers*. New York: Plenum Press.
- Arentoft, A., Byrd, D., Monzones, J., Coulehan, K., Fuentes, A., Rosario, A., ... Rivera Mindt, M. (2015). Socioeconomic status and neuropsychological functioning: Associations in an ethnically diverse HIV + cohort. *Clinical Neuropsychologist*, 29(2), 232–254. Available from <https://doi.org/10.1080/13854046.2015.1029974>.

- Arentoft, A., Byrd, D., Robbins, R. N., Monzones, J., Miranda, C., Rosario, A., . . . Rivera Mindt, M. (2012). Multidimensional effects of acculturation on English-language neuropsychological test performance among HIV + Caribbean Latinas/os. *Journal of Clinical and Experimental Neuropsychology*, 34(8), 814–825. Available from <https://doi.org/10.1080/13803395.2012.683856>.
- Barker-Collo, S. L. (2001). The 60-item Boston Naming Test: Cultural bias and possible adaptations for New Zealand. *Aphasiology*, 15(1), 85–92.
- Betancourt, H., & Lopez, S. R. (1993). The study of culture, ethnicity, and race in American psychology. *American Psychologist*, 48, 629–637.
- Boone, K. B., Victor, T. L., Wen, J., Razani, J., & Pontón, M. (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology*, 22(3), 355–365. Available from <https://doi.org/10.1016/j.acn.2007.01.010>.
- Buré-Reyes, A., Hidalgo-Ruzzante, N., Vilar-López, R., Gontier, J., Sánchez, L., Pérez-García, M., & Puente, A. E. (2013). Neuropsychological test performance of Spanish speakers: Is performance different across different Spanish-speaking subgroups? *Journal of Clinical and Experimental Neuropsychology*, 35(4), 404–412. Available from <https://doi.org/10.1080/13803395.2013.778232>.
- Byrd, D. A., Touradji, P., Tang, M. X., & Manly, J. J. (2004). Cancellation test performance in African American, Hispanic, and White elderly. *Journal of the International Neuropsychological Society*, 10(3), 401–411. Available from <https://doi.org/10.1017/S1355617704103081>.
- Candelaria, M. A., & Llorente, A. M. (2009). The assessment of the Hispanic child. In C. R. Reynolds, & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 401–427). New York, NY: Springer.
- Chiao, J. Y., & Blizinsky, K. D. (2010). Culture-gene coevolution of individualism–collectivism and the serotonin transporter gene. *Proceedings of Biological Sciences*, 277(1681), 529–537. Available from <https://doi.org/10.1098/rspb.2009.1650>.
- Church, T. A. (2016). Personality traits across cultures. *Current Opinion in Psychology*, 8, 22–30.
- Coffey, D. M., Marmol, L., Schock, L., & Adams, W. (2005). The influence of acculturation on the Wisconsin Card Sorting Test by Mexican Americans. *Archives of Clinical Neuropsychology*, 20(6), 795–803. Available from <https://doi.org/10.1016/j.acn.2005.04.009>.
- Colby, S. L., & Ortman, J. M. (2014). *Projections of the size and composition of the U.S. population: 2014 to 2060, current population reports, P25-1143*. Washington, DC: U.S. Census Bureau.
- Cosentino, S., Manly, J. J., & Mungas, D. (2007). Do reading tests measure the same construct in multiethnic and multilingual older persons? *Journal of the International Neuropsychological Society*, 13(2), 228–236. Available from <https://doi.org/10.1017/S1355617707070257>.
- Cruice, M. N., Worrall, L. E., & Hickson, L. M. H. (2000). Boston naming test results for healthy older Australians: A longitudinal and cross-sectional study. *Aphasiology*, 14, 143–155.
- Cuellar, I. (2000). Acculturation as 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.
- Culbertson, W. C., & Zilmer, E. A. (2001). *Tower of London—Drexel University (ToLDx) Technical manual*. NY: Multi-Health Systems.

- D'Elia, L. F., Satz, P., Uchiyama, C. L., & White, T. (1996). *Color trails test: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Dana, R. H. (1993). *Multicultural assessment perspectives for professional psychology*. Boston: Allyn and Bacon.
- Daugherty, J. C., Puente, A. E., Fasfous, A. F., Hidalgo-Ruzzante, N., & Pérez-García, M. (2017). Diagnostic mistakes of culturally diverse individuals when using North American neuropsychological tests. *Applied Neuropsychology: Adult*, 24(1), 16–22. Available from <https://doi.org/10.1080/23279095.2015.1036992>.
- Del Ser, T., González-Montalvo, J. I., Martínez-Espinosa, S., Delgado-Villapalos, C., & Bermejo, F. (1997). Estimation of premorbid intelligence in Spanish people with the Word accentuation test and its application to the diagnosis of dementia. *Brain and Cognition*, 33(3), 343–356. Available from <https://doi.org/10.1006/brcg.1997.0877>.
- Diez Roux, A. V., Merkin, S. S., Arnett, D., Chambless, L., Massing, M., Nieto, F. J., . . . . . Watson, R. L. (2001). Neighborhood of residence and incidence of coronary heart disease. *The New England Journal of Medicine*, 345(2), 99–106. Available from <https://doi.org/10.1056/NEJM200107123450205>.
- Dilworth-Anderson, P., Hendrie, H. C., Manly, J. J., Khachaturian, A. S., Fazio, S., & Social, Behavioral and Diversity Research Workgroup of the Alzheimers Association. (2008). Diagnosis and assessment of Alzheimer's disease in diverse populations. *Alzheimers and Dement*, 4(4), 305–309. Available from <https://doi.org/10.1016/j.jalz.2008.03.001>.
- Elbulok-Charcape, M. M., Rabin, L. A., Spadaccini, A. T., & Barr, W. B. (2014). Trends in the neuropsychological assessment of ethnic/racial minorities: A survey of clinical neuropsychologists in the United States and Canada. *Cultural Diversity and Ethnic Minority Psychology*, 20(3), 353–361. Available from <https://doi.org/10.1037/a0035023>.
- Fasfous, A. F., Hidalgo-Ruzzante, N., Vilar-López, R., Catena-Martínez, A., & Pérez-García, M. (2013). Cultural differences in neuropsychological abilities required to perform intelligence tasks. *Archives of Clinical Neuropsychology*, 28(8), 784–790. Available from <https://doi.org/10.1093/arclin/act074>.
- Fasfous, A. F., Puente, A. E., Pérez-Marfil, M. N., Cruz-Quintana, F., Peralta-Ramirez, I., & Pérez-García, M. (2013). Is the color trails culture free? *Archives of Clinical Neuropsychology*, 28(7), 743–749. Available from <https://doi.org/10.1093/arclin/act062>.
- Fyffe, D. C., Mukherjee, S., Barnes, L. L., Manly, J. J., Bennett, D. A., & Crane, P. K. (2011). Explaining differences in episodic memory performance among older African Americans and Whites: The roles of factors related to cognitive reserve and test bias. *Journal of the International Neuropsychological Society*, 17(4), 625–638. Available from <https://doi.org/10.1017/S1355617711000476>.
- Gasquoine, P. G. (2009). Race-norming of neuropsychological tests. *Neuropsychology Review*, 19(2), 250–262. Available from <https://doi.org/10.1007/s11065-009-9090-5>.
- Gasquoine, P. G., Croyle, K. L., Cavazos-Gonzalez, C., & Sandoval, O. (2007). Language of administration and neuropsychological test performance in neurologically intact Hispanic American bilingual adults. *Archives of Clinical Neuropsychology*, 22(8), 991–1001. Available from <https://doi.org/10.1016/j.acn.2007.08.003>.
- Glymour, M. M., & Manly, J. J. (2008). Lifecourse social conditions and racial and ethnic patterns of cognitive aging. *Neuropsychology Review*, 18(3), 223–254. Available from <https://doi.org/10.1007/s11065-008-9064-z>.
- Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Culture and psychology. Cross-cultural research methods in psychology* (pp. 46–74). New York: Cambridge University Press.

- Harris, J. G., & Llorente, A. M. (2005). Cultural consideration in the use of the Wechsler Intelligence Scale for Children—fourth edition (WISC-IV). In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical use and interpretation: Scientist-practitioner perspectives* (pp. 382–413). Burlington: Elsevier Academic.
- Harris, J. G., Tulsky, D. S., & Schultheis, M. T. (2003). Assessment of non-native English speaker: Assimilating history and research findings to guide clinical practice. In S. D. Tulsky, D. H. Saklofske, R. K. Heaton, R. Bornstein, M. F. Ledbetter, G. J. Chelune, R. J. Ivnik, & A. Prifitera (Eds.), *Clinical interpretation of the WAIS-III and WMS-III* (pp. 343–390). San Diego, CA: Academic Press.
- Ibanez-Casas, I., Daugherty, J.C., Leonard, B.E., Perez-Garcia, M., & Puente, A. (2017). Protocol for the development of a domain specific computerized battery for cross-cultural neurocognitive assessment: The EMBRACED Project. *Paper presented at the 45th annual meeting of the international neuropsychological society, New Orleans, LA.*
- International Test Commission. (2016). The ITC guidelines for translating and adapting tests (2nd ed.). Retrieved from [www.InTestCom.org](http://www.InTestCom.org).
- Ivanova, I., & Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta Psychologica (Amst)*, 127(2), 277–288. Available from <https://doi.org/10.1016/j.actpsy.2007.06.003>.
- Judd, T., Capetillo, D., Carrión-Baralt, J., Mármol, L. M., Miguel-Montes, L. S., Navarrete, M. G., . . . . . NAN Policy and Planning Committee. (2009). Professional considerations for improving the neuropsychological evaluation of Hispanics: A National Academy of Neuropsychology education paper. *Archives of Clinical Neuropsychology*, 24(2), 127–135. Available from <https://doi.org/10.1093/arclin/acp016>.
- Kaufman, J. S., Cooper, R. S., & McGee, D. L. (1997). Socioeconomic status and health in blacks and whites: The problem of residual confounding and the resilience of race. *Epidemiology*, 8, 621–628.
- Kennepohl, S. (1999). Toward a cultural neuropsychology: An alternative view and a preliminary model. *Brain and Cognition*, 41(3), 365–380. Available from <https://doi.org/10.1006/brcg.1999.1138>.
- Krahn, G. L., Walker, D. K., & Correa-De-Araujo, R. (2015). Persons with disabilities as an unrecognized health disparity population. *American Journal of Public Health*, 105 (Suppl 2), S198–S206. Available from <https://doi.org/10.2105/AJPH.2014.302182>.
- LaCalle, J. (1987). Forensic psychological evaluations through an interpreter: Legal and ethical issues. *American Journal of Forensic Psychology*, 5, 29–43.
- Llorente, A. M. (2007). *Principles of neuropsychological assessment with Hispanics: Theoretical Foundations and clinical practice*. New York: Springer.
- Llorente, A. M. (2008). *Principles of neuropsychological assessment with Hispanics: Theoretical foundations and clinical practice*. New York: Springer.
- Llorente, A. M., Taussig, M. I., Satz, P., & Perez, L. M. (2000). Trends in American immigration: Influences on neuropsychological assessment and inferences with ethnic minority population. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology*. New York: Kulwer Academic/Plenum Publishers.
- Lopez-Class, M., Gonzalez Castro, F., & Ramirez, A. G. (2011). Conceptions of acculturation: A review and statement of critical issues. *Social Science & Medicine*, 72, 1555–1562.
- Lucas, J. A., Ivnik, R. J., Smith, G. E., Ferman, T. J., Willis, F. B., Petersen, R. C., & Graff-Radford, N. R. (2005). Mayo's Older African Americans normative studies: Norms for Boston naming test, controlled oral word association, category fluency, animal naming,

- token test, WRAT-3 reading, trail making test, stroop test, and judgment of line orientation. *The Clinical Neuropsychologist*, 19(2), 243–269. Available from <https://doi.org/10.1080/13854040590945337>.
- Luria, A. R. (1976). *Cognitive development: Its cultural and social foundations*. Cambridge, MA: Harvard University Press.
- Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed). New York: Basic Books.
- Maj, M., D'Elia, L., Satz, P., Janssen, R., Zaudig, M., Uchiyama, C., ... World Health Organization, Division of Mental Health/Global Programme on AIDS. (1993). Evaluation of two new neuropsychological tests designed to minimize cultural bias in the assessment of HIV-1 seropositive persons: a WHO study. *Archives of Clinical Neuropsychology*, 8(2), 123–135.
- Manly, J. J. (2006). Deconstructing race and ethnicity: Implications for measurement of health outcomes. *Medical Care*, 44(11 Suppl 3), S10–S16. Available from <https://doi.org/10.1097/01.mlr.0000245427.22788.be>.
- Manly, J. J., Byrd, D. A., Touradji, P., & Stern, Y. (2004). Acculturation, reading level, and neuropsychological test performance among African American elders. *Applied Neuropsychology*, 11(1), 37–46. Available from [https://doi.org/10.1207/s15324826an1101\\_5](https://doi.org/10.1207/s15324826an1101_5).
- Manly, J. J., & Echemendia, R. J. (2007). Race-specific norms: using the model of hypertension to understand issues of race, culture, and education in neuropsychology. *Archives of Clinical Neuropsychology*, 22(3), 319–325. Available from <https://doi.org/10.1016/j.acn.2007.01.006>.
- Manly, J. J., Jacobs, D. M., Touradji, P., Small, S. A., & Stern, Y. (2002). Reading level attenuates differences in neuropsychological test performance between African American and White elders. *Journal of the International Neuropsychological Society*, 8 (3), 341–348.
- Manly, J. J., Schupf, N., Tang, M. X., & Stern, Y. (2005). Cognitive decline and literacy among ethnically diverse elders. *Journal of Geriatric Psychiatry and Neurology*, 18(4), 213–217. Available from <https://doi.org/10.1177/0891988705281868>.
- Marin, G. (1992). Issues in the measurement of acculturation among Hispanics. In K. F. Geisinger (Ed.), *The psychological testing of Hispanics* (pp. 235–251). Washington, DC: American Psychological Association.
- Marin, G., & Gamba, R. J. (1996). A new measurement of acculturation for Hispanics: The bidimensional acculturation scale for Hispanics (BAS). *Hispanic Journal of Behavioral Sciences*, 18, 297–316.
- McCrae, R. R., & Terracciano, A. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology*, 89(3), 407–425. Available from <https://doi.org/10.1037/0022-3514.89.3.407>.
- Melendez, F. (2001). Forensic assessment of Hispanics. In M. O. Ponton, & J. Leon-Carrion (Eds.), *Neuropsychology and the Hispanic patient: A clinical handbook* (pp. 321–340). Mahwah, NJ: Lawrence Erlbaum and Associates.
- Miranda, C., Arce Rentería, M., Fuentes, A., Coulehan, K., Arentoft, A., Byrd, D., ... Rivera Mindt, M. (2016). The relative utility of three english language dominance measures in predicting the neuropsychological performance of HIV + Bilingual Latino/a adults [Formula: see text]. *The Clinical Neuropsychologist*, 30(2), 185–200. Available from <https://doi.org/10.1080/13854046.2016.1139185>.
- Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed). Oxford: Oxford University Press.

- Mungas, D., Reed, B. R., Farias, S. T., & Decarli, C. (2009). Age and education effects on relationships of cognitive test scores with brain structure in demographically diverse older persons. *Psychology and Aging, 24*(1), 116–128. Available from <https://doi.org/10.1037/a0013421>.
- Mungas, D., Reed, B. R., Haan, M. N., & González, H. (2005). Spanish and English neuropsychological assessment scales: relationship to demographics, language, cognition, and independent function. *Neuropsychology, 19*(4), 466–475. Available from <https://doi.org/10.1037/0894-4105.19.4.466>.
- Nell, V. (2000). *Cross-cultural neuropsychological assessment: Theory and practice*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- Noble, J. M., Manly, J. J., Schupf, N., Tang, M. X., & Luchsinger, J. A. (2012). Type 2 diabetes and ethnic disparities in cognitive impairment. *Ethnicity & Disease, 22*(1), 38–44.
- Noble, K. G., McCandliss, B. D., & Farah, M. J. (2007). Socioeconomic gradients predict individual differences in neurocognitive abilities. *Developmental Science, 10*(4), 464–480. Available from <https://doi.org/10.1111/j.1467-7687.2007.00600.x>.
- Norman, M. A., Moore, D. J., Taylor, M., Franklin, D., Cysique, L., Ake, C., ... Group, H. (2011). Demographically corrected norms for African Americans and Caucasians on the Hopkins verbal learning test-revised, brief visuospatial memory test-revised, stroop color and word test, and wisconsin card sorting test 64-card version. *Journal of Clinical and Experimental Neuropsychology, 33*(7), 793–804. Available from <https://doi.org/10.1080/13803395.2011.559157>.
- Ostrosky-Solis, F., Ardila, A., & Rosselli, M. (1998). Neuropsychological test performance in illiterate subjects. *Archives of Clinical Neuropsychology, 3*(17), 645–660.
- Ostrosky-Solis, F., & Ramirez, M. (2004). Effects of culture and education on neuropsychological testing: A preliminary study with indigenous and nonindigenous population. *Applied Neuropsychology, 11*, 186–193.
- Paige, L. E., Ksander, J. C., Johndro, H. A., & Gutchess, A. H. (2017). Cross-cultural differences in the neural correlates of specific and general recognition. *Cortex, 91*, 250–261. Available from <https://doi.org/10.1016/j.cortex.2017.01.018>.
- Paradis, M. (2008). Bilingualism and neuropsychiatric disorders. *Journal of Neurolinguistics, 21*, 199–230.
- Park, D. C., & Huang, C. M. (2010). Culture wires the brain: A cognitive neuroscience perspective. *Perspectives on Psychological Science, 5*(4), 391–400. Available from <https://doi.org/10.1177/1745691610374591>.
- Perez-Arce, P. (1999). The influence of culture on cognition. *Archives of Clinical Neuropsychology, 14*(7), 581–592.
- Ponton, M. (2001). Research and assessment issues with Hispanic populations. In M. Ponton, & J. Leon-Carrion (Eds.), *Neuropsychology and the Hispanic patient: A clinical handbook* (pp. 39–58). Mahwah, NJ, USA: Erlbaum.
- Ponton, M., & Corona-Lomonaco, M. E. (2007). Cross-cultural issues in neuropsychology: Assessment of Hispanic patient. In B. P. Uzzell, M. Ponton, & A. Ardila (Eds.), *International handbook of cross-cultural neuropsychology* (pp. 265–283). New York, NY: Routledge.
- Puente, A. E., Perez-Garcia, M., Vilar-Lopez, R., Hidalgo-Ruzzante, N. A., & Fasfous, A. (2013). Neuropsychological assessment of culturally and educationally dissimilar individuals. In F. A. Paniagua, & A.-M. Yamada (Eds.), *Handbook of multicultural mental health: Assessment and treatment of diverse population* (pp. 225–242). San Diego, CA: Academic Press.

- Razani, J., Burciaga, J., Madore, M., & Wong, J. (2007). Effects of acculturation on tests of attention and information processing in an ethnically diverse group. *Archives of Clinical Neuropsychology*, 22(3), 333–341. Available from <https://doi.org/10.1016/j.acn.2007.01.008>.
- Razani, J., Murcia, G., Tabares, J., & Wong, J. (2006). The effects of culture on WASI test performance in ethnically diverse individuals. *The Clinical Neuropsychologist*, 21, 776–788.
- Rivera Mindt, M., Arentoft, A., Kubo Germano, K., D'Aquila, E., Scheiner, D., Pizzirusso, M., ... Gollan, T. H. (2008). Neuropsychological, cognitive, and theoretical considerations for evaluation of bilingual individuals. *Neuropsychology Review*, 18(3), 255–268. Available from <https://doi.org/10.1007/s11065-008-9069-7>.
- Rivera Mindt, M., Byrd, D., Ryan, E. L., Robbins, R., Monzones, J., Arentoft, A., ... Morgello, S. (2008). Characterization and sociocultural predictors of neuropsychological test performance in HIV + Hispanic individuals. *Cultural Diversity and Ethnic Minority Psychology*, 14, 315–325.
- Rivera Mindt, M., Byrd, D., Saez, P., & Manly, J. J. (2010). Increasing culturally competent neuropsychological services for ethnic minority populations: a call to action. *The Clinical Neuropsychologist*, 24(3), 429–453. Available from <https://doi.org/10.1080/13854040903058960>.
- Robertson, K., Liner, J., & Heaton, R. (2009). Neuropsychological assessment of HIV-infected populations in international settings. *Neuropsychology Review*, 19(2), 232–249. Available from <https://doi.org/10.1007/s11065-009-9096-z>.
- Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition*, 52(3), 326–333.
- Rosselli, M., Ardila, A., Salvatierra, J., Marquez, M., Matos, L., & Weekes, V. A. (2002). A cross-linguistic comparison of verbal fluency tests. *International Journal of Neuroscience*, 112(6), 759–776.
- Ryan, E. L., Baird, R., Mindt, M. R., Byrd, D., Monzones, J., Bank, S. M., & Bank, M. H. B. (2005). Neuropsychological impairment in racial/ethnic minorities with HIV infection and low literacy levels: Effects of education and reading level in participant characterization. *Journal of the International Neuropsychological Society*, 11(7), 889–898.
- Saez, P. A., Bender, H. A., Barr, W. B., Rivera Mindt, M., Morrison, C. E., Hassenstab, J., ... Vazquez, B. (2014). The impact of education and acculturation on nonverbal neuropsychological test performance among Latino/a patients with epilepsy. *Applied Neuropsychology: Adult*, 21(2), 108–119. Available from <https://doi.org/10.1080/09084282.2013.768996>.
- Savignon, S. J. (2007). Beyond communicative language teaching: What's ahead? *Journal of Pragmatics*, 39, 207–220.
- Sifers, S. K., Puddy, R. W., Warren, J. S., & Roberts, M. C. (2002). Reporting of demographics, methodology, and ethical procedures in journals in pediatric and child psychology. *Journal of Pediatric Psychology*, 27(1), 19–25.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance on African Americans. *Journal of Personality and Social Psychology*, 69, 797–811.
- Stephenson, M. (2005). Development and validation of the Stephenson Multigroup Acculturation Scale (SMAS). *Psychological Assessment*, 12, 77–88.
- Sternberg, R. J., & Grigorenko, E. L. (2004). Intelligence and culture: How culture shapes what intelligence means, and the implications for a science of well-being. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1449), 1427–1434. Available from <https://doi.org/10.1098/rstb.2004.1514>.

- Touradji, P., Manly, J. J., Jacobs, D. M., & Stern, Y. (2001). Neuropsychological test performance: A study of non-Hispanic white elderly. *Journal of Clinical and Experimental Neuropsychology*, 23, 643–649.
- Unger, J. B., Gallaher, P., Shakib, S., Ritt-Olson, A., Palmer, P. H., & Johnson, A. C. (2002). The AHIMSA acculturation scale: A new measure of acculturation for adolescents in a multicultural society. *Journal of Early Adolescence*, 22, 225–251.
- United Nations, Department of Economic and Social Affairs, Population Division. (2016). International Migration Report 2015 (ST/ESA/SER.A/384).
- van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.
- Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.
- Way, B. M., & Lieberman, M. D. (2010). Is there a genetic contribution to cultural differences? Collectivism, individualism and genetic markers of social sensitivity. *Social Cognitive and Affective Neuroscience*, 5(2–3), 203–211. Available from <https://doi.org/10.1093/scan/nsq059>.
- Woodcock, R. W., & Munoz-Sandoval, A. F. (1996). *Bateria Wodcock-Munoz: Pruebas de habilidad cognitiva-revisada, supplemental manual*. Chicago: Riverside Publishing Company.
- Zebrowski, C. M., Vega, M., & Llorente, A. M. (2015). Cultural and linguistic issues in the assessment and treatment of pediatric cancer survivors. In G. Mucci, & L. Torno (Eds.), *Handbook of long term care of the childhood cancer survivor. Specialty topics in pediatric neuropsychology* (pp. 299–313). Boston, MA: Springer.

# Technological developments in assessment

20

*Robert L. Kane<sup>1</sup> and Thomas D. Parsons<sup>2,3,4</sup>*

<sup>1</sup>Cognitive Consults and Technology LLC, Washington Neuropsychology Research Group, Neuropsychology Associates of Fairfax, Fairfax, VA, United States, <sup>2</sup>Director, NetDragon Digital Research Centre, University of North Texas, Denton, TX, United States, <sup>3</sup>Director, Computational Neuropsychology and Simulation (CNS) Lab, Department of Psychology, University of North Texas, Denton, TX, United States, <sup>4</sup>Professor, College of Information, University of North Texas, Denton, TX, United States

## Introduction

In medicine and related fields, examinations are in essence assessments or interrogations of functional systems. A complete physical examination may include assessing the integrity of biological and organ systems whose functioning is critical for a patient's health. In psychology, this review may involve cognitive systems relating to areas such as affective self-regulation, language, memory, various components of executive functioning, aspects of attention, and problem-solving. Psychologists also assess behavioral and symptom patterns that relate to different clinical presentations. Data obtained from the review of systems can be used to establish whether or not an individual is showing problems within one or more cognitive and affective domains, may help inform diagnostic considerations, be used to assess capacity related to occupational or legal considerations, and provide information related to treatment planning. In neuropsychology, the once-important objective of lesion localization has become less relevant with advances in neuroimaging (Bigler, 1991; Baxendale & Thompson, 2010). However, assessing the cognitive status of an individual, as well as the integrity of various cognitive and affective systems, remains an important psychological contribution to patient evaluation, and traditional assessment methods and tests have functioned admirably in making psychology a valued part of modern health care. While it is important to acknowledge the success of psychological assessment to date, no profession can afford to stand still and in an era marked by rapid advances in technology in almost every area of life, it is critical to address to what degree technology can enhance the psychological examination and potentially expand uses for cognitive assessment. The goal of this chapter is to provide an overview of the current and potential impact of technology on cognitive assessment and to provide a roadmap for future developments making use of technology to enhance the assessment and rehabilitation process. An underlying theme is that technology should be used not because it is there but because it can

introduce capabilities and options that at minimum enhance efficiency and more importantly augment the assessment and interpretation processes. For this chapter, we have identified seven main areas where technology can have a substantial impact on the practice of psychological assessment. These include: (1) enhancing the efficiency and reliability of the assessment process; (2) expanding the types of behaviors that can be assessed, including the implementation of scenario-based assessment; (3) increasing access to care; (4) improving assessment models by linking cognitive domains and biological systems; (5) refining diagnosis and prediction using analytics; (6) augmenting cognitive rehabilitation and self-monitoring; and (7) expanding research methods.

### ***Assessment: enhancing efficiency and reliability***

Once the microcomputer became available in the 1970s it was inevitable that psychologists would begin exploring the potential of this device for both research and clinical assessment. Much of the early work was accomplished by researchers addressing specific problems or by individual clinicians whose goal was to develop tests for specific purposes. For example, [Acker and Acker \(1982\)](#) in England developed a brief battery that could be used to assess individuals with alcohol problems. Their battery ran on an Apple II computer. [Branconnier \(1986\)](#) developed The Alzheimer's Disease Assessment Battery. It also ran on the Apple II and was constructed so that both the patient and the examiner worked on separate computer terminals. For general clinical assessment, [Swiercinsky \(1984\)](#) developed the SAINT-II. This system was composed of 10 tests, many of which were based on the author's experience with the Halstead-Reitan Test Battery. The SAINT-II included tests assessing spatial orientation, motor persistence, verbal and visual memory, visual search speed, vocabulary, sequencing, rhythm discrimination, numerical skill, and set shifting ([Kane & Reeves, 1997](#)).

A great deal of early work took place in laboratories run by psychologists in the Department of Defense. These individuals were engaged in human performance research and were interested in expanding the tests available for cognitive assessment and in developing batteries where test measures could be repeated in order to assess the effects of various environmental stressors. Much of this early work took place outside the awareness of most psychologists and publications were often in the form of technical reports. The fact that much of this early work was not visible or easily discoverable by most psychologists led [Kane and Kay \(1992\)](#) to publish a review of the work that had been accomplished in the area of computerized testing at that point in time. In this review they discussed basic issues in computerized assessment, outlined and categorized the types of tasks that had been implemented by different test batteries, and reviewed available batteries with respect to technical considerations, psychometric properties, and supporting research. Since that time there has been an expansion in both the use and availability of computerized test measures.

### *Advantages and challenges in early adoption*

Dating from the first attempts to implement computers as cognitive assessment devices, there was an appreciation of the potential advantages as well as the challenges and cautions that surrounded computerized assessment. The advantages included standardization in test administration superior to that of the human examiner, scoring accuracy, the ability to integrate response timing into a variety of tasks in order to better assess processing speed and efficiency, expanded test metrics that capture variability in performance, the ability to integrate adaptive testing for both test and item selection, and the ability to incorporate tests not easily done with booklets or pieces of paper. The cautions included that computers could be deceptive with respect to timing accuracy and developers of automated tests had to be aware of issues related to response input methods, operating systems, program coding, and computer architecture that could affect response timing and the measurement of test performance. In the early years of automated testing, computers were relatively new devices not yet fully integrated into daily life. Hence, there were also concerns about how an individual would react to being tested on a computer. These concerns were supported by studies addressing the effects of computer familiarity on test performance (Johnson & Mihal, 1973; Iverson, Brooks, Ashton, Johnson, & Gualtieri, 2009). The biggest limitation in adopting computerized testing for clinical assessment was that available technology did not permit the assessment of important language-based skills, including verbal memory. Algorithms for analyzing speech patterns for indications for anxiety, depression, or unusual syntax or word usage were also not available or were in the early stages of development. Constraints related to speech recognition and language analysis limited the types of tests and methods that have been implemented on computer for clinical assessment.

### *Technological advances*

While early adopters of computerized testing faced some challenges, it is unlikely that these limitations will persist for much longer. There are two thrusts in technology that will have a substantial impact on the ability to implement clinically relevant tasks via computer. These are speech (and language) recognition and adaptive computational models (i.e., artificially intelligent algorithms). Computational models will be discussed later as they relate to clinical decision making. However, improvements in the ability of automated systems to recognize spoken language will dramatically affect how computers can be integrated into cognitive and behavioral assessment. At present, it takes a human examiner to assess verbal memory, confrontation naming, generative fluency, and expository speech. Advances in speech and language recognition will eventually make it possible to assess a full range of cognitive and affective functions using automated methods. Also, it is likely that by implementing artificial neural networks, the computer will be able to pick up additional data informative to the assessment process (Mitchell & Xu, 2015). Humans are skilled at assessing emotional pauses, facial expression, and vocal intonations. Computers continue to have limitations for assessing mood and

affect. However, the gap between humans and computers in assessing emotion, while wide, continues to narrow through development and implementation of artificial neural network architectures constructed to handle the fusion of different modalities (facial features, prosody and lexical content in speech; [Fragapanagos & Taylor, 2005](#); [Wu, Parsons, & Narayanan, 2010](#)). While early computational approaches to speech recognition and language processing focused on automated analyses of linguistic structures and the development of lab-based technologies (machine translation, speech recognition, and speech synthesis), current approaches can be implemented in real-world speech-to-speech translation engines that can identify sentiment and emotion ([Hirschberg & Manning, 2015](#)).

As language processing improves, and as feature detection becomes more sophisticated, the range of clinically relevant aspects of how patients present will continue to expand in ways that will both enhance and challenge current approaches to psychological assessment. Studies are already underway to assess temporal characteristics of spontaneous speech in acquired brain injury and neurodegenerative disease ([Sztaloczki, Hoffmann, Vincze, Kalman, & Pakaski, 2015](#)). Computer-automated language analysis may help the early diagnosis of cognitive decline. The temporal characteristics of spontaneous speech (e.g., speech tempo, number and length of pauses in speech) are markers of cognitive disorders ([Roark, Mitchell, Hosom, Hollingshead, & Kaye, 2011](#)).

Future capabilities aside, there are a number of more immediate advantages to computerized testing that have yet to be fully exploited. The ability of computers to permit the implementation of tasks not well done or in some cases not possible through traditional means will be discussed later in this chapter. However, there are a number of other ways in which the computer can enhance and make more efficient the testing process. Booklet based tests are inherently clumsy. Test materials fray, making the process of presenting stimuli at exact intervals even more unreliable by compounding normal human error. Booklet based testing is inherently inefficient. It requires a constant shifting of materials along with hand scoring. Scores are later tabulated and either entered into a scoring program or require the examiner to refer to tables in order to record standard scores. Even assuming that no scoring error has taken place, a substantial amount of examination time is taken up in what are essentially administrative activates. Current standards allow for this time to be billed as time spent testing and scoring. If present health care trends persist, professional activities will be reimbursed based on outcomes and proven utility and not time spent. Depending on how changes in health care reimbursement affect psychology, failing to embrace technology as a way of modernizing the assessment process may well have economic consequences.

## **Expanding tasks and scenario-based assessment**

Task-based assessment is focused on having the patient demonstrate skills in specific cognitive domains. While most measures require more than one skill to

complete, specific tests attempt to target a specific skill. In life, individuals must implement a number of skills and integrate a number of abilities to manage demands in their environment. Some traditional measures such as the Wisconsin Card Sorting Test (WCST; [Heaton, 2003](#)) and Part B of the Trail Making Test ([Reitan, 1955](#)) are used by psychologists because of the degree of cognitive flexibility needed to complete the task. Furthermore, tests like the WCST and the Halstead Category Test (HCT; [Reitan & Wolfson, 1985](#)) are used because of their apparent ability to assess problem-solving based on feedback.

### ***Computer-automated assessment of multitasking***

Traditional tests do not capture divided attention and how well an individual is able to allocate cognitive resources when having to tackle different tasks during a specific period of time. Early on, test developers appreciated that computers could be used to implement complex tasks in addition to those used in traditional assessment approaches such as the WCST and the HCT. Developers also appreciated that computers could be used to implement actual tests of divided attention where the person being tested would have to allocate resources between two or more tests running concurrently. The Complex Cognitive Assessment Battery (CCAB; [Kane & Kay, 1992](#)) was developed by DoD as a group of tasks that involved a number of complex problem-solving skills. While no longer available, the CCAB was a model for using computers to expand the assessment of complex reasoning skills. SynWork, developed by Tim Elsmore ([Kane & Kay, 1992](#)), was an early demonstration of a true divided attention or multi tasking test. SynWork presented up to four tasks that ran in different quadrants of the computer screen. Scores were produced for each task individually and there was an overall score that captured how well an individual was able to allocate their cognitive resources among tasks. The motivation for SynWork (synthetic work) was that day-to-day demands rely on an integration of skills and having the ability to effectively switch one's focus between and among competing demands. The total score produced by SynWork increased as the test taker effectively completed a task and was reduced when the test taker failed to respond correctly to a task within a given timeframe, or needed to request reminders. SynWork, while attempting to model day-to-day demands, was task based. The person taking the test performed different cognitive tasks that were presented simultaneously rather than individually.

### ***Virtual environments for ecologically valid assessments***

There is an apparent need for psychology to expand beyond its current conceptual and experimental frameworks. Burgess and colleagues ([Burgess et al., 2006](#)) argued that most cognitive assessments in use today fail to represent the actual functional capacities inherent in cognitive (e.g., executive) functions. These authors suggest that traditional psychological assessments like the WCST assess a hypothetical cognitive construct that can be inferred from research findings (e.g., correlation

between two variables). Cognitive construct measures like the WCST were not originally designed to be used as clinical measures (Burgess et al., 2006). Instead, these measures were useful tools for cognitive assessment in normal populations which later found their way into the clinical realm to aide in assessing constructs that are important to carrying out real-world activities. Goldstein (1996) questioned this approach because it is difficult to ascertain the extent to which performance on measures of basic constructs translates to functional capacities within the varying environments found in the real world. Psychologists need assessments that further our understanding about the ways in which the brain enables persons to interact with their environment and organize everyday activities.

Virtual environments (VEs) are increasingly considered as potential aids in enhancing the ecological validity of psychological assessments (Campbell et al., 2009; Renison, Ponsford, Testa, Richardson, & Brownfield, 2012; Parsons, 2015). This increased interest is at least in part due to recent enhancements in 3D rendering capabilities that accelerated graphics considerably and allowed for greatly improved texture and shading in computer graphics. Earlier virtual reality equipment suffered a number of limitations, such as being large and unwieldy, difficult to operate, and very expensive to develop and maintain. Over the past decade, researchers have steadily progressed in making VE hardware and software more reliable, cost effective, and acceptable in terms of size and appearance (Bohil, Alicea, & Biocca, 2011). Today VEs offer advanced computer interfaces that allow patients to become immersed within a computer-generated simulation of everyday activities. As with any test there is need for validation of these measures (see Parsons, McMahan, & Kane, *in press*; Parsey & Schmitter-Edgecombe, 2013).

Like other computerized automated psychological assessments, VEs have greater computational capacities that allow for enhanced administration: reliable and controlled stimulus presentation, automated response logging, database development, and data analytic processing. Given that VEs allow for precise presentation and control of dynamic perceptual stimuli, psychologists can use them to provide assessments that combine the veridical control and rigor of laboratory measures with a verisimilitude that reflects real life situations (Parsons, 2015). Additionally, the enhanced computation power allows for increased accuracy in the recording of cognitive and emotional responses in a perceptual environment that systematically presents complex stimuli. VE-based psychological assessments can provide a balance between naturalistic observation and the need for exacting control over key variables. In sum, there is a growing body of evidence that supports the position that VE-based psychological assessments allow for real-time evaluation of multifarious cognitive and affective responses in order to measure complex sets of skills and behaviors that may more closely resemble real-world functional abilities (see Bohil et al., 2011; Kane & Parsons, 2017; Parsey & Schmitter-Edgecombe, 2013, Parsons, 2016, 2017; Parsons & Phillips, 2016; Parsons, Carlew, Magtoto, & Stonecipher, 2017, Parsons, Gaglioli, & Riva, 2017).

## Access to care and telehealth

Three key barriers to receiving psychological assessment are geographic distance from a qualified provider, the cost of assessment, and wait times between requesting an appointment and actually being seen. The growth of telehealth in medicine has far outpaced the exploration of remote assessment in psychology. According to the website eVisit ([eVisit, 2016](#)), in the year 2018, 7 million patients will receive services through telemedicine and as of August 2015, 29 states require health insurers to pay for telemedicine services. Unfortunately, clinical neuropsychology has fallen way behind in addressing this need.

A new medium for delivering psychological assessments has emerged as a result of the Internet. Recent surveys have revealed that over 3.1 billion people now have access to the Internet. The distribution of this number by country reveals the following: China = 642 million; United States = 280 million; India = 243 million; Japan = 109 million; Brazil = 108 million; Russia = 84 million, among others ([Stats, 2015](#)). In the United States 86.75% of residents have access to the Internet. Telemedicine is an area that has developed for the use and exchange of medical information from one site to another via electronic communications, information technology, and telecommunications. When researchers are discussing “telemedicine,” they typically mean synchronous (interactive) technologies such as videoconferencing or telephony to deliver patient care. When the clinical services involve mental health or psychiatric services, the terms “telemental health” and “telepsychiatry” are generally used ([Yellowlees, Shore, & Roberts, 2010](#)).

### ***Remote psychological assessment***

Remote psychological assessment is a recent development in telemedicine, in which psychologists administer remotely behavioral and cognitive assessments to expand the availability of specialty services ([Cullum & Grosch, 2012](#)). Evaluation of the patient is performed via a personal computer, digital tablet, smartphone, or other digital interface to administer, score, and aide interpretation of these assessments ([Cullum, Hynan, Grosch, Parikh, & Weiner, 2014](#)). Preliminary evaluation of patient acceptance of this methodology has revealed that it appears to be well accepted by consumers. For example, in the area of cognitive assessment, Parikh and colleagues ([Parikh et al., 2013](#)) found 98% satisfaction and approximately two-thirds of participants reported no preference between assessment via video teleconferencing and traditional in-person assessment.

Remote behavioral assessment is often done by way of interview and may include the administration of short questionnaires to assess pertinent symptoms. Remote cognitive assessment is just beginning to develop and for many the concept seems challenging at best. However, four models have emerged for remote cognitive assessment that have the potential to increase access to care for patients and potentially reduce costs including that for travel. Model 1 is a minor variation of employing a technician for test administration. It involves the interview being done

remotely by a psychologist with tests administered by a technician collocated with the patient. While statistics are not available, this method likely represents the current, or at least most frequent, implementation of remote cognitive assessment. In Model 2, both the clinical interview and test administration are accomplished remotely. This model has some limitations—it may require some tests to be renormed, and may also involve an assistant to help the patient sign on and set up certain test materials. Nevertheless, research done to date supports the viability of this model for both short screening tests such as the Mini Mental State Examination (Loh, Ramesh, Maher, Saligari, Flicker, & Goldswain, 2004; Loh, Donaldson, Flicker, Maher, & Goldswain, 2007; McEachern, Kirk, Morgan, Crossley, & Henry, 2008) as well as for more extensive cognitive assessment batteries (Cullum, Weiner, Gehrmann, & Hynan, 2006; Cullum et al., 2014; Jacobsen, Sprenger, Andersson, & Krogstad, 2002). The attractiveness of this model is that it addresses the reality that a trained technician may not always be available at sites distant from the location of the examining psychologist. Model 3 takes advantage of the fact that there are a number of computerized tests that can be set up for remote administration and that require minimal verbal input and guidance from an examiner. Tests can be downloaded to run locally on the computer used by the person taking the test, with data securely transferred back to the examiner. In some cases tests can be Internet-based. A recently published pilot study demonstrated the viability of this model using the Automated Neuropsychological Assessment Metrics system (ANAM; Settle, Robinson, Kane, Maloni, & Wallin, 2015). This study compared test scores when patients with multiple sclerosis (MS) were assessed in person, in a different hospital room from the examiner, and at home. Results from the study demonstrated that test results were comparable when the same patients were tested remotely in different locations to those obtained with traditional in-person test administration. To preserve timing accuracy cognitive tests ran locally on the patient's computer while the examiner monitored and communicated with the patient remotely. Model 4 is essentially a hybrid model that acknowledges that different approaches to remote cognitive assessment can be combined when assessing patients who are not collocated with the examining psychologist.

Computer-based tests have expanded access to care by permitting data to be obtained from various groups of individuals potentially at risk for injury. A subset of the ANAM battery (Reeves, Winter, Bleiberg, & Kane, 2007) was implemented by NASA as the Spaceflight Cognitive Assessment Tool for Windows (WinSCAT; Kane, Short, Sipes, & Flynn, 2005). To date, WinSCAT has been used on 47 expeditions to the International Space Station (K.A. Seaton, personal communication, May 16, 2016). As a result of concerns about brain injury occurring during combat, as part of the 2008 National Defense Authorization Act (Congress, 2008), Congress mandated baseline testing on all deploying Service members. As of this writing baseline testing has been obtained on 1,140,445 Service members using a subset of ANAM tests. The database for individuals tested includes over 2 million assessments (D. Marion, personal communication, May 16, 2016). The ability to test individuals in space along with the ability to obtain baseline and post injury information on large numbers of individuals performing in hazardous environments

was possible only through using technology to expand models for cognitive assessment. The ImPACT test system (<https://impacttest.com/research/>) has been used, along with other computerized test systems, to gather baseline and post injury data on athletes. While these uses have been selective and focused on specific populations, they are also models for obtaining and storing data that may be useful throughout the life span when assessing the effects of disease or injury. These systems have made possible the concept of making cognition an additional medical endpoint for longitudinal health monitoring. Test instruments used for longitudinal health monitoring should be carefully developed and validated.

## **Linking cognitive domains and biological systems**

If a psychological evaluation involves the systematic exploration of pertinent cognitive and affective systems, then it is incumbent to define these systems and to determine specific aspects of these systems that should be assessed. Current domains and subdomains have emerged through studies in cognitive psychology, the examination of patients with different lesions and pathologies, and though factor analytic studies. Larrabee (2014) has argued for the development of an ability-focused battery based on factor analytic studies in order to produce a research-based approach to the systematic investigation of clinically relevant cognitive domains. While there is substantial merit to his approach, there is also an argument for expanding Halstead's (1947) original goal of trying to capture biological intelligence. That is, it is important to understand what the brain is actually doing, how does it process information, how are neural networks organized, and what is the best way to measure this behaviorally. An example of this approach is that used by Posner (2011, 2016) (see also Posner & Rothhart, 2007) to help define the structure of attention networks and to tie behavioral tasks into these networks.

### ***Neuroimaging***

Neuroimaging has gained widespread use in clinical research and practice. As a result, some objectives previously found within the expertise of psychology—principally lesion localization and laterality of function—have been almost completely replaced by neuroimaging. While neuroimaging has taken advantage of advances in computerization and neuroinformatics, psychological assessments have not kept pace with advances in neuroscience and reflect nosological attempts at classification that occurred prior to contemporary neuroimaging. This lack of development in psychological assessment makes it very difficult to develop clinical psychological models. The call for advances in psychological assessment is not new. Twenty years ago Dodrill (1997) argued that psychological assessments by clinical psychologists had made much less progress than would be expected in both absolute terms and in comparison with the progress made in other clinical neurosciences. When one compares progress in psychological assessment with progress in assessments

found in clinical neurology, it is apparent that while the difference may not have been that great prior the appearance of computerized tomographic scanning (in the 1970s), advances since then (e.g., magnetic resonance imaging) have given clinical neurologists a dramatic edge. Neuroimaging with its rapidly increasing capabilities will continue to play a role in our understanding of the functional organization of the brain. Technological advances in neuroimaging of brain structure and function offer great potential for revolutionizing psychological assessment (Bilder, 2011). Bigler (1991) has also made a strong case for integrating cognitive assessment with neuroimaging as a more potent method for understanding brain pathology and its effects. This integration can be aided with the use of computerized metrics time locked to imaging sequences.

### ***Advancing innovative neurotechnologies***

The National Institute of Mental Health's Research Domain Criteria (RDoC) is a research framework for developing and implementing novel approaches to the study of mental disorders. The RDoC integrates multiple levels of information to enhance understanding of basic dimensions of functioning underlying the full range of human behavior from normal to abnormal. The RDoC framework consists of functional constructs (i.e., concepts that represent a specified functional behavior dimension) categorized in aggregate by the genes, molecules, circuits, etc. used to measure it. In turn, the constructs are grouped into higher level domains of functioning that reflect current knowledge of the major systems of cognition, affect, motivation, and social behavior (see Insel et al., 2010).

The BRAIN Initiative represents an ambitious but achievable set of goals for advances in science and technology. Since the announcement of the BRAIN Initiative, dozens of leading academic institutions, scientists, technology firms, and other important contributors to neuroscience have responded to this call. A group of prominent neuroscientists have developed a 12-year research strategy for the National Institutes of Health to achieve the goals of the initiative. The BRAIN Initiative may do for neuroscience what the Human Genome Project did for genomics. It supports the development and application of innovative technologies to enhance our understanding of brain function. Moreover, the BRAIN initiative endeavors to aid researchers in uncovering the mysteries of brain disorders [e.g., Alzheimer's, Parkinson's, depression, and traumatic brain injury (TBI)]. It is believed that the initiative will accelerate the development and application of new technologies for producing dynamic imaging of the brain that express the ways in which individual brain cells and complex neural circuits interact at the speed of thought.

### ***Enhancing diagnosis and behavioral prediction: computational assessment/neuropsychology***

As noted above, two major thrusts in computer-based technology are language recognition and artificial intelligence (AI). There are different approaches to AI

including developing decision algorithms, deep machine learning, and training neural networks. Algorithms are typically developed by humans who review research data and develop decision rules. Machine learning lets the computer analyze data for patterns some of which may prove important though not be immediately intuitive. Neural networks require training to discern patterns important for whatever determination is to be made. In all cases, the better and more abundant the data, the better the classification or prediction. In contributing to the diagnostic process, the psychologist looks for patterns that conform to the known literature about various clinical conditions. By-and-large, as a profession, psychologists do a good job at this. However, the psychologist's ability to refine pattern detection can be enhanced through the accumulation and analysis of large data sets. The challenge in doing this is more organizational than technical. Computers have been able to defeat champions playing the quiz show Jeopardy and at strategic games like Chess and Go. Technologies for accumulating and analyzing data and exploring rules and relationships exist and are becoming increasingly powerful and available. It is now up to the profession of psychology to pool resources and to take full advantage of data analytics to bring about advances in defining disease patterns and in making predictions regarding the ecological consequences of cognitive and behavioral test performance.

## Cognitive rehabilitating and self-monitoring

Psychologists spend a substantial amount of time characterizing performance patterns including identifying deficit areas. There have also been substantial efforts over the years to find ways to adjust and/or train cognitive skills. Data supporting cognitive training has often been disappointing. There have been positive results noted, although there has been a trend for effect sizes to lessen when improvements in performance in single arm studies are measured against those obtained in studies using control groups. Despite the difficulty of the task, clinicians and researchers continue to explore ways to train and/or rehabilitate cognitive processes. It is unlikely that technology will produce an easy fix with respect to developing approaches to cognitive training. However, there are data to indicate that technology can offer helpful tools and approaches for enhancing cognitive processes.

The jury is still out regarding the efficacy of commercially available Internet-based cognitive enrichment programs ([Simons et al., 2016](#)). In 2014 a group of leading cognitive psychologists and neuroscientist were brought together at the Stanford Center on Longevity and the Berlin Max Planck Institute for Human Development to assesses the research pertaining to the use of brain games for cognitive enhancement ([Longevity, 2014](#)). The group concluded that more research was required to judge the effectiveness of these programs and couched its recommendations in terms of opportunity costs. That is, if one had other meaningful ways to remain engaged then they felt the data supporting online cognitive enhancement programs were insufficient to pull back from these other activities to focus on

computer training. Absent other ways of remaining active and stimulated, then engaging in online cognitive activates might be a reasonable way to spend time. Other scientists have gone on record supporting the efficacy of cognitive rehabilitation ([www.cognitivetrainingdata.org](http://www.cognitivetrainingdata.org)). The [Simons et al. \(2016\)](#) review concluded there was extensive evidence that brain-training interventions enhances performance on the tasks used for training, less evidence that people get better when performing closely related tasks, and little evidence to support generalization to distantly related tasks or day-to-day functioning. The fundamental issue seems to be the effectiveness of training rather than the method of delivery. The Internet has the capability of bringing interventions to more people once the efficacy of the rehabilitation approach has been established.

### ***Computers for cognitive training***

There are a number of approaches to using computers for cognitive training. Reviewing each of these is beyond the scope of this chapter. Two programs will be mentioned here as examples of systematic efforts to start from data based theory, develop a computer-based approach based on theory and previous data, and then undertake the rigorous work of attempting to validate the method. Recently Chiaravalloti and her colleagues published randomized control trials of a method called the modified story memory technique to improve memory in patients with MS and TBI ([Chiaravalloti, Moore, Nikelshpur, & DeLuca, 2013](#); [Chiaravalloti, Dobryakova, Wylie, & DeLuca, 2015](#)). These researchers were able to define the nature of the memory problem experienced by both MS and TBI patients as primarily that of acquiring information. They developed a computer-based method guided by previous studies to teach acquisition strategies and then validated the method through randomized control studies. Another example is a program being developed by [Chen and colleagues \(2011\)](#) that uses a video game like presentation of various scenarios to provide guided experimental learning and to teach performance enhancing techniques. In addition to being research-based, the approach by Chen et al. was designed from its initial stages to focus on the transfer of skills to the real world and to be adaptable to telehealth. While a number of rehabilitation programs and exercises have been implemented on the computer, the use of well-crafted scenarios designed to work hand-in-hand with structured training offers the potential to substantially enhance rehabilitation methods. Developing these methods for telehealth will result in more patients being offered these services.

### ***Smartphones for psychological assessment***

Smartphones offer psychologists mobile computing capabilities and given their mobility and ubiquity in the general population they offer new options for research in cognitive science ([Dufau et al., 2011](#)). [Brouillette and colleagues \(2013\)](#) developed a new application that utilizes touch screen technology to assess attention and processing speed. Initial validation was completed using an elderly nondemented

population. Findings revealed that their color shape test was a reliable and valid tool for the assessment processing speed and attention in the elderly. These findings support the potential of smartphone-based assessment batteries for attentional processing in geriatric cohorts. From a mental health perspective, smartphone applications have been developed for the assessment and treatment of various conditions including depression, anxiety, substance use, sleep disturbance, suicidality, psychosis, eating disorders, stress, and gambling (Donker et al., 2013). While supporting data are limited, initial findings indicate potential uses for these apps for addressing behavioral conditions (Donker et al., 2013). One expects the development of smartphone apps in mental health to continue in view of their potential to extend interventions beyond the therapist's office and to reinforce patient's engagement in their treatment.

### ***Ecological momentary assessments***

Psychologists are often interested in the everyday real-world behavior of their patients because brain injury and its functional impairments are expressed in real-world contexts. An unfortunate limitation is that many psychological assessments do little to tap into activities of daily living, quality of life, affective processing, and life stressors. These aspects of the patient's life are surveyed using global, summary, or retrospective self-reports. The prominence of global questionnaires can keep psychologists from observing and studying dynamic fluctuations in behavior over time and across situations. Further, these questionnaires may obfuscate the ways in which a patient's behavior varies and is governed by context. In reaction to the frequent reliance of psychologists on global, retrospective reports (and the serious limits they place on accurately characterizing, understanding, and changing behavior in real-world settings), some psychologists are turning to Ecological Momentary Assessment (EMA; Cain, Depp, & Jeste, 2009; Waters & Li, 2008; Waters et al., 2014; Schuster, Mermelstein, & Hedeker, 2015). EMA is characterized by a series of (often computer and/or smartphone-based) repeated assessments of cognitive, affective (including physiological), and contextual experiences of participants as they take part in everyday activities (Shiffman, Stone, & Hufford, 2008; Jones & Johnston, 2011).

EMA uses modern methods to capture performance and behavioral characteristics in the course of everyday functioning. It makes possible capturing information in more depth and in broader contexts than would be obtained from naturalistic observation alone. Useful data can also be gathered from directly observing individuals performing everyday tasks. Some observational approaches are task-specific (e.g., Goverover, Chiaravalloti, & DeLuca, 2010; Goverover & DeLuca, 2015; Goverover, Chiaravalloti, & DeLuca, 2015) and others involve capturing broader samples of behavior as with the Multiple Errands Task (Dawson et al., 2009). However, EMA can be used to capture a range of behaviors in multiple environments without the observer having to be collocated with the patient or subject.

## Expanding research options

Some early developments in computerized testing were driven and shaped by research needs. Within the DoD there was a need to expand testing capabilities to better study the effects of environmental stressors on performance and to study medication side effects (Kane & Kay, 1992). This implementation of cognitive testing meant that individuals being evaluated would have to be tested on multiple occasions, under different conditions, and that tests and methods would have to be implemented for repeated administrations. The microcomputer was an obvious tool to develop tests designed for repeated measures assessment and to implement test measures potentially sensitive to various factors that might affect human performance. The need to assess groups of people at risk was also behind the development of the Neurobehavioral Evaluation System 2 (NES 2; Letz & Baker, 1988). Tests in the NES 2—like those developed by DoD—were designed for repeated measures assessment, and in addition were consistent with recommendations made by the World Health Organization for assessing the effects of environmental toxins (Kane & Kay, 1992). Another test battery that set the stage for the use of computerized tests in pharmaceutical studies was the Memory Assessment Clinics battery (Larrabee & Crook, 1988). This computerized test system used technology that existed during that time period (CD-ROM) to present tests that attempted to mirror everyday tasks such as telephone dialing, name–face association, learning a grocery list, and remembering in which room the test taker placed various objects, among other measures. Currently, computerized tests designed for repeated measures are used extensively in pharmaceutical research. In addition, computer driven technologies are permitting the measurement of critical performance skills in ecologically relevant ways. Driving safety is an important concern when evaluating the effects and side effects of some medications. A recent development has been to employ driving simulators in pharmaceutical studies. For example, Kay and Feldman (2013) reported the results of a study that employed a desktop computer-based simulator to investigate whether the use of armodafinil could improve driving performance in patients with obstructive sleep apnea who demonstrated excessive daytime sleepiness. The use of simulators and virtual reality (VR) scenarios to augment current approaches in cognitive and behavioral research is likely to experience growth in the coming years. Clinical outcomes in dementia include not only the measurement of basic cognitive skills but also an assessment of day-to-day functioning. Everyday skills are typically measured via rating scales or tests that combine knowledge-based questions with a limited sampling of skills. Using validated VR scenarios to measure functional behavior is likely the next logical step in making use of advances in technology to expand assessment capabilities in psychological research. Using scenario-based assessment in this context has the potential advantage of producing functional metrics that are not dependent on subjective report.

Two challenges in conducting successful clinical trials are those of recruitment and retention of subjects. Challenges to retention may involve subjects moving or finding returning to a research facility expensive, difficult, or just tiresome. Having

the ability to accomplish at least some aspects of follow-up in a person's home will likely reduce subject dropout in research investigations. Remote psychological assessment was discussed earlier in this chapter with regard to its clinical implications. There is also a research role for remote cognitive assessment as a method for following subjects for longer periods of time despite obstacles imposed by travel. The technologies needed for this capability (e.g., Internet connection, cell phone) are increasingly ubiquitous and technologies exist for secure communication. The implementation of a remote assessment system will need to include a method of verifying that the intended responder is the person answering questions or performing tasks.

## Conclusions

In this chapter, we discussed the ability of technology to both enhance and expand current approaches to psychological assessment. In some cases, potential contributions of technology involve streamlining and making more efficient current assessment models. In other cases, technology can bring about paradigm shifts in how we conceptualize and measure cognitive processes, design interventions, and expand the reach of psychological services. The area of psychological assessment has been criticized for the sluggishness with which it embraces change (Sternberg, 1997). In 1997 Dodrill (1997) noted that psychologists had made much less progress in the way in which we as a profession approach assessment than would be expected in absolute terms and in comparison with other clinical neurosciences. In 1987 Meehl (1987) commented on how perplexing it would be if clinical psychologists lagged behind professions such as medicine and investment analysis—not to mention functions such as controlling operations in factories—in making full use of the power of the computer. The underlying theme of this chapter is that technology needs to be embraced by psychology not because it is there or because it enhances our credibility in the 21st century, but because it substantially expands our capabilities to assess and treat patients and to engage in research. To date, clinical psychologists have underused technology. This underuse has in part been resistance and in part been a function of the fact that those developing technology have not placed sufficient emphasis on clinical relevance. For clinical psychology to continue to prosper, both of these impediments need to be overcome.

## References

- Acker, W., & Acker, C. (1982). *Bexley Maudsley automated psychological screening and Bexley Maudsley category sorting test: Manual*. Windsor, Great Britain: NFER-Nelson.
- Baxendale, S., & Thompson, P. (2010). Beyond localization: The role of traditional neuropsychological tests in an age of imaging. *Epilepsia*, 51(11), 2225–2230.

- Bigler, E. D. (1991). Neuropsychological assessment, neuroimaging, and clinical neuropsychology: A synthesis. *Archives of Clinical Neuropsychology*, 6(3), 113–132.
- Bilder, R. M. (2011). Neuropsychology 3.0: Evidenced-based science and practice. *Journal of the International Neuropsychological Society*, 17(1), 7–13.
- Bohil, C. J., Alicea, B., & Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *National Review of Neurosciences*, 12(12), 752–762.
- Branconnier, R. J. (1986). A computerized battery for behavioral assessment in Alzheimer's disease. In L. W. Poon, T. Cook, K. L. Davis, et al. (Eds.), *Handbook for clinical memory assessment of older adults* (pp. 189–196). Washington, D.C.: American Psychological Association.
- Brouillette, R. M., Foil, H., Fontenot, S., Correro, A., Allen, R., Martin, C. K., ... Keller, J. N. (2013). Feasibility, reliability, and validity of a smartphone base application for the assessment of cognitive function in the elderly. *PLoS One*, 8(6), e65925.
- Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L. M., Dawson, D. R., ... Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, 12(2), 194–209.
- Cain, A. E., Depp, C. A., & Jeste, D. V. (2009). Ecological momentary assessment in aging research: A critical review. *Journal of Psychiatric Research*, 43, 987–996.
- Campbell, Z., Zakzanis, K. K., Jovanovski, D., Joordens, S., Mraz, R., & Graham, S. J. (2009). Utilising virtual reality to improve the ecological validity of clinical neuropsychology: An fMRI case study elucidating the neural basis of planning by comparing the Tower of London with a three-dimensional navigation task. *Applied Neuropsychology*, 16, 295–306.
- Chen, A. J.-W., Novakovic-Agopian, T., Nycum, T. J., Song, S., Turner, G. R., Hills, N. K., ... D'Esposito, M. (2011). Training of goal-directed attention regulation enhances control over neural processing for individuals with brain injury. *Brain*, 134(Pt 5), 1541–1554.
- Chiavallotti, N. D., Dobryakova, E., Wylie, G. R., & DeLuca, J. (2015). Examining the efficacy of the modified story memory technique (mSMT) in persons with TBI using functional magnetic resonance imaging (fMRI): The TBI-MEM trial. *Journal of Head Trauma Rehabilitation*, 30(4), 261–269.
- Chiavallotti, N. D., Moore, N. B., Nikelshpur, O. M., & DeLuca, J. (2013). An RCT to treat learning impairment in multiple sclerosis: The MEMREHAB trial. *Neurology*, 81(24), 2066–2072.
- Congress, U. S. (2008). National Defense Authorization Act: 4986; Section 1618. <<https://www.govtrack.us/congress/bills/110/hr4986>>.
- Cullum, C. M., & Grosch, M. G. (2012). Teleneuropsychology. In K. Myers, & C. Turvey (Eds.), *Telemental health: Clinical, technical, and administrative foundations for evidence-based practice* (pp. 275–294). Amsterdam: Elsevier.
- Cullum, C. M., Hynan, L. S., Grosch, M., Parikh, M., & Weiner, M. F. (2014). Teleneuropsychology: Evidence for video conference-based neuropsychological assessment. *Journal of the International Neuropsychological Society*, 20, 1–6.
- Cullum, C. M., Weiner, M. F., Gehrmann, H. R., & Hynan, L. S. (2006). Feasibility of telecognitive assessment in dementia. *Assessment*, 13(4), 385–390.
- Dawson, D. R., Anderson, N. D., Burgess, P., Cooper, E., Krpan, K. M., & Stuss, D. T. (2009). Further development of the Multiple Errands Test: Standardization, scoring, reliability, and ecological validity for the Baycrest version. *Archives of Physical Medicine and Rehabilitation*, 90(11), S41–S51.

- Dodrill, C. B. (1997). Myths of neuropsychology. *The Clinical Neuropsychologist*, 11, 1–17.
- Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M. R., & Christensen, H. (2013). Smartphones for smarter delivery of mental health programs: A systematic review. *Journal of Medical Internet Research*, 15(11), e247.
- Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X., ... Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS One*, 6(9), e24974.
- eVisit (2016). 36 Telemedicine statistics you should know. Retrieved from <https://evisit.com/36-telemedicine-statistics-know/>.
- Fragapanagos, N., & Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Networks*, 18(4), 389–405.
- Goldstein, G. (1996). Functional considerations in neuropsychology. In R. J. Sbordone, & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 75–89). Delray Beach, FL: GR Press/St. Lucie Press.
- Goverover, Y., Chiaravalloti, N. D., & DeLuca, J. (2010). Actual reality: A new approach to functional assessment in persons with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation*, 91, 252–260.
- Goverover, Y., & DeLuca, J. (2015). Actual reality: Using the internet to assess everyday functioning after traumatic brain injury, E-pub *Brain Injury*. Available from <https://doi.org/10.3109/02699052.2015.1004744>.
- Goverover, Y., Chiaravalloti, N., & DeLuca, J. (2015). Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS) and performance of everyday life tasks: Actual reality, E-pub; 2016 *Multiple Sclerosis Journal*, 22(4), 544–550. Available from <https://doi.org/10.1177/1352458515593637>.
- Halstead, W. C. (1947). *Brain and intelligence: A quantitative study of the frontal lobes*. Chicago, IL: University of Chicago Press.
- Heaton, R. K. (2003). *Wisconsin card sorting test computer version 4.0*. Odessa, FL: Psychological Assessment Resources.
- Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261–266.
- Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7), 748–751.
- Iverson, G. L., Brooks, B. L., Ashton, V. L., Johnson, L. G., & Gaultier, C. T. (2009). Does familiarity with computers affect computerized neuropsychological test performance. *Journal of Clinical and Experimental Neuropsychology*, 31, 594–604.
- Jacobsen, S. E., Sprenger, T., Andersson, S., & Krogstad, J. M. (2002). Neuropsychological assessment and telemedicine: A preliminary study examining the reliability of neuropsychological services performed via telecommunication. *Journal of the International Neuropsychological Society*, 9, 472–478.
- Johnson, J. H., & Mihal, W. L. (1973). The performance of blacks and whites in computerized versus manual testing environments. *American Psychologist*, 28, 694–699.
- Jones, M., & Johnston, D. (2011). Understanding phenomena in the real world: The case for real time data collection in health services research. *Journal of Health Services Research & Policy*, 16, 172–176.
- Kane, R. L., & Kay, G. G. (1992). Computerized assessment in neuropsychology: A review of tests and test batteries. *Neuropsychology Review*, 3(1), 1–117.
- Kane, R. L., & Reeves, D. L. (1997). Computerized test batteries. In A. M. Horton, D. Wedding, & J. Webster (Eds.), *The neuropsychology handbook* (Vol. 1). New York, NY: Springer.

- Kane, R. L., Short, P., Sipes, W., & Flynn, C. F. (2005). Development and validation of the spaceflight cognitive assessment tool for Windows (WinSCAT). *Aviation, Space, and Environmental Medicine*, 76(6, Suppl.), B183–B191.
- Kane, R. L., & Parsons, T. D. (Eds.). (2017). *The role of technology in clinical neuropsychology*. Oxford: Oxford University Press.
- Kay, G. G., & Feldman, N. (2013). Effects of Armodafinil on simulated driving and self-report measures in obstructive sleep apnea patients prior to treatment with continuous positive airway pressure. *Journal of Clinical Sleep Medicine*, 9(5), 445–454.
- Larrabee, G. J. (2014). Test validity and performance validity: Considerations in providing a framework for development of an ability-focused neuropsychological test battery. *Archives of Clinical Neuropsychology*, 29, 695–714.
- Larrabee, G. J., & Crook, T. (1988). A computerized everyday memory battery for assessing treatment effects. *Psychopharmacology Bulletin*, 24(4), 695–697.
- Letz, R., & Baker, E. L. (1988). *Neurobehavioral evaluation system 2: Users manual (version 4.2)*. Winchester, MA: Neurobehavioral Systems.
- Loh, P. K., Donaldson, M., Flicker, L., Maher, S., & Goldswain, P. (2007). Development of a telemedicine protocol for the diagnosis of Alzheimer's disease. *Journal of Telemedicine and Telecare*, 13(2), 90–94.
- Loh, P. K., Ramesh, P., Maher, S., Saligari, J., Flicker, L., & Goldswain, P. (2004). Can patients with dementia be assessed at a distance? The use of telehealth and standardized assessments. *Internal Medicine Journal*, 34, 239–242.
- Longevity SCO. (2014) A consensus of the brain training industry from the scientific community.
- McEachern, W., Kirk, A., Morgan, D. G., Crossley, M., & Henry, C. (2008). Reliability of the MMSE administered in-person and by telehealth. *Canadian Journal of Neurological Sciences*, 35(5), 643–646.
- Meehl, P. E. (1987). Theory and practice: Reflections of an academic clinician. In E. F. Bourg, R. J. Bent, J. E. Callan, N. F. Jones, J. McHolland, & G. Stricker (Eds.), *Standards and evaluation in the education and training of professional psychologists: Knowledge, attitudes, and skills* (pp. 7–23). Transcript Press.
- Mitchell, R. L., & Xu, Y. (2015). What is the value of embedding artificial emotional prosody in human–computer interactions. *Frontiers in Psychology*, 6.
- Parikh, M., Grosch, M. C., Graham, L. L., Hynan, L. S., Weiner, M., Shore, J. H., ... Cullum, C. M. (2013). Consumer acceptability of brief videoconference-based neuropsychological assessment in older individuals with and without cognitive impairment. *The Clinical Neuropsychologist*, 27(5), 808–817.
- Parsey, C. M., & Schmitter-Edgecombe, M. (2013). Applications of technology in neuropsychology. *The Clinical Neuropsychologist*, 27(8), 1328–1361.
- Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in clinical, affective, and social neurosciences. *Frontiers in Human Neuroscience*, 9, 1–19.
- Parsons, T. D. (2016). *Clinical neuropsychology and technology: What's new and how we can use it*. New York, NY: Springer Press.
- Parsons, T. D., & Phillips, A. (2016). Virtual reality for psychological assessment in clinical practice. *Practice Innovations*, 1, 197–217.
- Parsons, T. D., Carlew, A. R., Magtoto, J., & Stonecipher, K. (2017). The potential of function-led virtual environments for ecologically valid measures of executive function in experimental and clinical neuropsychology. *Neuropsychological Rehabilitation*, 37(5), 777–807.

- Parsons, T. D., Gaglioli, A., & Riva, G. (2017). Virtual environments in social neuroscience. *Brain Sciences*, 7(42), 1–21.
- Parsons, T. D. (2017). *Cyberpsychology and the brain: The interaction of neuroscience and affective computing*. Cambridge: Cambridge University Press.
- Parsons, T. D., McMahan, T., & Kane, R. (in press). Practice parameters facilitating adoption of advanced technologies for enhancing neuropsychological assessment paradigms. *The Clinical Neuropsychologist*.
- Posner, M. I. (2011). *Attention in a social world*. Oxford: Oxford University Press.
- Posner, M. I. (2016). Orienting of attention: then and now. *The Quarterly Journal of Experimental Psychology*, 69, 1864–1875.
- Posner, M. I., & Rothbart, M. K. (2007). Research on attention networks as a model for the integration of psychological science. *Annual Review of Psychology*, 58, 1–23.
- Reeves, D. L., Winter, K., Bleiberg, J., & Kane, R. L. (2007). ANAM genogram: Historical perspectives, description, and current endeavors. *Archives of Clinical Neuropsychology*, 22S, S15–S37.
- Reitan, R. M. (1955). The relation of the trail making test to organic brain damage. *Journal of Consulting Psychology*, 195, 393–394.
- Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.
- Renison, B., Ponsford, J., Testa, R., Richardson, B., & Brownfield, K. (2012). The ecological and construct validity of a newly developed measure of executive function: The virtual library task. *Journal of the International Neuropsychological Society*, 18, 440–450.
- Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2081–2090.
- Schuster, R. M., Mermelstein, R. J., & Hedeker, D. (2015). Acceptability and feasibility of a visual working memory task in an ecological momentary assessment paradigm. *Psychological Assessment*, 27(4), 1463.
- Settle, J. R., Robinson, S. A., Kane, R., Maloni, H. W., & Wallin, M. T. (2015). Remote cognitive assessments for patients with multiple sclerosis: A feasibility study. *Multiple Sclerosis Journal*, 21, 1072–1079. Available from <https://doi.org/10.1177/1352458514559296>.
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32.
- Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., ... Stine-Morrow, E. A. (2016). Do “brain-training” programs work? *Psychological Science in the Public Interest*, 17(3), 103–186.
- Stats, I. L. (2015). Retrieved from <<http://www.internetlivestats.com/internet-users-by-country/>> Accessed 28.05.15.
- Sternberg, R. J. (1997). Intelligence and lifelong learning: What's new and how can we use it? *American Psychologist*, 52(10), 1134.
- Swiercinsky, D. (1984). *Computerized neuropsychological assessment*. Houston, TX: International Neuropsychological Society.
- Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 7, 195.
- Waters, A. J., & Li, Y. (2008). Evaluating the utility of administering a reaction time task in an ecological momentary assessment study. *Psychopharmacology*, 197, 25–35.

- Waters, A. J., Szeto, E. H., Wetter, D. W., Cinciripini, P. M., Robinson, J. D., & Li, Y. (2014). Cognition and craving during smoking cessation: An ecological momentary assessments study. *Nicotine & Tobacco Research*, 16(Suppl. 2), S111–S118.
- Wu, D., Parsons, T. D., & Narayanan, S. S. (2010). Acoustic feature analysis in speech emotion primitives estimation. *Interspeech*, 785–788.
- Yellowlees, P., Shore, J., & Roberts, L. (2010). Practice guidelines for videoconferencing-based telemental health-October 2009. *Telemedicine and e-Health*, 16(10), 1074–1089.

# Index

*Note:* Page numbers followed by “f” and “t” refer to figures and tables, respectively.

## A

- AACN. *See* American Academy of Clinical Neuropsychology (AACN)  
AAN. *See* American Academy of Neurology (AAN)  
AAS. *See* Addiction acknowledgment scale (AAS)  
Ability Profile, The, 181  
Abstract attitude, 228  
Academic Focus scale, 175  
Accommodation, 216, 555–556  
Acculturation, 211, 558–559  
Accuracy, 55, 159–160, 209, 232, 470<sup>t</sup>  
Achievement testing, 143–144, 152<sup>t</sup>, 162–163  
    in clinical practice, 151  
    in public schools, 149–150  
Achilles’ Heel of Outcome Research in Behavior Therapy, 17  
ACID profile, 75–76  
ACS. *See* Advanced Clinical Solutions (ACS)  
ACSI-28. *See* Athletic Coping Skills Inventory-28 (ACSI-28)  
ACT. *See* American College Testing (ACT)  
ACTFL. *See* American Council on the Teaching of Foreign Languages (ACTFL)  
Activation factor, 366  
“Active imagination”, 419–420  
AD. *See* Alzheimer’s disease (AD)  
ADAS. *See* Alzheimer’s Disease Assessment Scale—Cognitive Subscale (ADAS)  
Addiction acknowledgment scale (AAS), 408<sup>t</sup>  
Addiction potential scale (APS), 408<sup>t</sup>  
Adequate testing of hypotheses, 204  
Adequate yearly progress (AYP), 150

- ADHD. *See* Attention deficit hyperactivity disorder (ADHD)  
ADIS-C/P. *See* Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions (ADIS-C/P)  
ADIS-IV. *See* Anxiety Disorders Interview Schedule for Children (ADIS-IV)  
ADOS. *See* Autism Diagnostic Observation Schedule (ADOS)  
Adult(s)  
    affective symptoms for, 369–375  
    Bech–Rafaelsen Mania Scale, 375  
    CDSS, 373–374  
    HRSD, 371–372  
    IDS, 372–373  
    YMRS, 374–375  
behavioral assessment with, 461–496  
comprehensive neuropsychological assessment, 227  
comprehensive batteries, 241–262  
problems in construction and standardization, 232–241  
development, 174  
diagnostic interviews, 356–362  
SES, 562  
symptom and behavior rating scales, 362–380  
Advanced Clinical Solutions (ACS), 122  
    social cognition, 129–130  
    suboptimal effort, 130  
        Word Choice subtest, 130  
Advancing innovative neurotechnologies, 582  
Adventuring Orientation, 174–175  
Affect Naming, 129  
Affective/physiological responses, 472–473, 475  
Affordable Care Act, 553

- Agency for Healthcare Research and Quality (AHRQ), 513
- Aggregation of information, 397
- Agility, 275
- Aging, 505, 515  
normal, 506–508
- Agnosia, 230–231
- AHRQ. *See* Agency for Healthcare Research and Quality (AHRQ)
- AI. *See* Artificial intelligence (AI)
- Alcohol Expectancy Questionnaire, 485
- ALD. *See* Arithmetic learning disorder (ALD)
- “Alleged psychodynamics”, 14
- ALSUP. *See* Assessment of Lagging Skills and Unsolved Problems (ALSUP)
- Alternate-form coefficients, 56
- Alzheimer’s disease (AD), 237, 505, 519
- Alzheimer’s Disease Assessment Battery, The, 574
- Alzheimer’s Disease Assessment Scale—Cognitive Subscale (ADAS), 514
- Alzheimer’s Disease-related Quality of Life scale (QoL-AD), 507–508
- AMA. *See* American Psychiatric Association (AMA)
- Ambiguous stimuli, 422, 461–462
- Ambulatory biosensor assessment, 487–488
- American Academy of Clinical Neuropsychology (AACN), 214, 554
- American Academy of Neurology (AAN), 512–513
- American College Testing (ACT), 144–145
- American Council on the Teaching of Foreign Languages (ACTFL), 560–561
- American Psychiatric Association (AMA), 538–539
- American Psychiatric Association (APA), 5, 20, 276–277, 398, 410, 513, 541, 543, 554
- APA Ethical Code, 543
- American Version of the National Adult Reading Test (AMNART), 132, 514
- AMI. *See* Auditory Memory Index (AMI)
- Analog behavioral observation method, 482–483
- Analysis of variance (ANOVA), 57
- Analyzing (Orientation Scale), 174
- ANAM. *See* Automated Neuropsychological Assessment Metrics (ANAM)
- Anatomical-organization interpretation model, 199
- Anchoring-and-adjustment bias, 219
- Anger scale (ANG), 409t
- Animal Naming test, 519
- ANOVA. *See* Analysis of variance (ANOVA)
- Antipsychotic medication, 363–364
- Antisocial practices scale (ASP), 409t
- ANX scale. *See* Anxiety scale (ANX scale)
- Anxiety, 447  
physiological indications of, 472–473
- Anxiety disorders, 34, 39, 345–346, 369–370
- Anxiety Disorders Interview Schedule for Children (ADIS-IV), 445
- Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions* (ADIS-C/P), 345–346
- Anxiety scale (ANX scale), 409t
- APA. *See* American Psychiatric Association (APA)
- Aphasia, 231  
Reitan Aphasia Screening Test, 246  
testing, 240
- Apple II computer, 574
- APS. *See* Addiction potential scale (APS)
- Aptitude testing, 143–144  
in 21st century, 162–163
- Aptitude treatment interaction approach (ATI approach), 86–87, 92–93
- Arithmetic learning disorder (ALD), 161–162
- Arithmetic subtest, 84–85, 115–116
- Armed Services Vocational Aptitude Battery (ASVAB), 148
- Army alpha test, 68, 104–105
- Artificial intelligence (AI), 65–66, 582–583
- Artificially intelligent algorithms, 575–576
- ASP. *See* Antisocial practices scale (ASP)
- Assessment approaches, 16–18, 203, 419, 461, 471  
in sports  
assessment of concussion, 288–289  
basics, 287–288  
constructs and behaviors, 280–281  
coping skills, 284–285

- disciplines of sports psychology, 276–277  
emotion regulation, 283–284  
mental toughness, 286  
neuropsychology, 287  
personality, 282–283  
psychology, 275, 278–280  
resilience, 285  
sport milieu, 277–278  
team cohesiveness, 286–287  
theory and goals, 201–209  
tools in forensic psychology, 539–540
- Assessment of Lagging Skills and Unsolved Problems (ALSUP), 449–450
- Assessment scales, 173
- Assimilation, 558–559
- ASVAB. *See* Armed Services Vocational Aptitude Battery (ASVAB)
- Asynchronous biological development, 196–197
- Athletic Coping Skills Inventory-28 (ACSI-28), 284
- ATI approach. *See* Aptitude treatment interaction approach (ATI approach)
- Attention, 516–517
- Attention deficit hyperactivity disorder (ADHD), 41, 75–76, 81, 151–154, 209, 216, 342, 435–436
- Attractiveness index, 43
- Auditory agnosia, 230–231
- Auditory attentional deficits, 230–231
- Auditory Memory, 124–125
- Logical Memory, 125
- Verbal Paired Associates, 125
- Auditory Memory Index (AMI), 124
- Autism Diagnostic Interview—Revised, 356–357
- Autism Diagnostic Observation Schedule (ADOS), 11
- Automated Neuropsychological Assessment Metrics (ANAM), 293, 579–580
- Average Impairment Rating, 244–247
- Avoidance Orientation Scale, 284–285
- AYP. *See* Adequate yearly progress (AYP)
- B**
- BADLS. *See* Bristol Activities of Daily Living Scale (BADLS)
- BAI. *See* Beck Anxiety Inventory (BAI)
- BASC-2. *See* Behavior Assessment System for Children, Second Edition (BASC-2)
- Basic Achievement Skills Individual Screener (BASIS), 151
- Basic Interest Scales (BIS), 178–179
- Bayesian methodology, 204–205
- BCSE. *See* Brief Cognitive Status Exam (BCSE)
- BD. *See* Block Design (BD)
- BDI-II. *See* Beck Depression Inventory (BDI-II)
- Bech–Rafaelsen Mania Scale (MAS), 375
- Beck Anxiety Inventory (BAI), 398, 464
- Beck Depression Inventory (BDI-II), 398
- Beck–Rafaelsen Scale, 371
- Bedwetting, 438
- Behavior assessment, 4, 13–19, 472, 477–488, 478*t*, 493–496
- with adults in clinical settings, 461–496
- ambulatory biosensor assessment, 487–488
- analog behavioral observation, 482–483
- behavioral observation, 480–481
- of children, 435
- of contexts, 449–453
- covert processes, 446–449
- history, 437–440
- overt behavior, 441–445
- conceptual foundations, 465–472
- constructing functional analysis and FACCD, 492–493
- DSMs and, 18–19
- EMA, 484–485
- functional behavioral interviews and questionnaires, 485–486
- identifying and evaluating causal relations, 490–491
- interventions effects based or not based on functional analysis, 496
- naturalistic behavioral observation, 481–482
- operationally defining target behaviors and causal variables, 489–490
- psychophysiological assessment, 486–487
- self-monitoring, 483–484
- Behavior Assessment System for Children, Second Edition (BASC-2), 47
- Behavior Rating of Executive Function (BRIEF), 444

- Behavior ratings and checklists, 443–444  
Behavior/behavioral, 461  
    abnormalities, 513  
    behavior-sampling strategy, 481–482  
    behavior–environment interactions, 467t  
    domains, 442  
    observation technique, 279, 397–398,  
        480–481  
    prediction, 582–583  
    rating scales, 362–380  
    therapy, 13, 15  
    variables, 50  
Behaviorism, 437–438  
Behaviorists, 438  
Bender–Gestalt test, 12–13, 229–230, 518  
Bennett Mechanical Comprehension Test  
    (BMCT), 147  
Berlin Max Planck Institute for Human  
    Development, 583–584  
BIA. *See* Brief Intellectual Ability (BIA)  
Bilingual test norms, 561  
Bilingualism, 559–560  
Binet–Simon Scale, 67–68, 104  
“Biological Intelligence” theory, 242  
Biological systems, 581–582  
Bipolar scales, 179  
BIS. *See* Basic Interest Scales (BIS)  
Bizarre Mentation scale (BIZ scale), 409t  
‘Black box’ approach, 293–294  
Block Design (BD), 85, 113–114, 518  
    subtest of WAIS-IV, 518  
    subtest of Wechsler scales, 230  
Bloom’s taxonomy, 37–39  
BMCT. *See* Bennett Mechanical  
    Comprehension Test (BMCT)  
BNSS. *See* Brief Negative Symptom Scale  
    (BNSS)  
Boston Diagnostic Aphasia Examination,  
    231–232  
Boston Naming Test, 242, 246  
BPRS. *See* Brief Psychiatric Rating Scale  
    (BPRS)  
Bradykinesia, 520  
Bradyphrenia, 520  
Brain  
    brain-behavior relationships, 203  
    brain-damaged patients, 229–230, 232  
    damage, 227–228  
    dysfunction, 202  
    functioning, 211  
    injury, 75–76, 215  
    and intelligence, 242  
    Wechsler–Bellevue studies of brain  
        lesion lateralization, 236  
BRAIN Initiative, 582  
Breadth, 435–436  
BRIEF. *See* Behavior Rating of Executive  
    Function (BRIEF)  
Brief Cognitive Status Exam (BCSE),  
    127–128  
Brief Intellectual Ability (BIA), 87–88  
Brief Negative Symptom Scale (BNSS), 377,  
    379–380  
Brief Psychiatric Rating Scale (BPRS), 7,  
    365–366, 366t  
Bristol Activities of Daily Living Scale  
    (BADLS), 507–508
- C**
- C1–C11 scales (clinical scales), 256  
CAFU. *See* Caregiver Assessment of  
    Function and Upset (CAFU)  
CAI. *See* Career Assessment Inventory (CAI)  
Calgary Depression Scale for Schizophrenia  
    (CDSS), 371, 373–374  
California Psychological Inventory (CPI),  
    398  
California Verbal Learning Test, 519  
Campbell Interest and Skill Survey (CISS),  
    170, 173–176  
    item pool and profile for, 174  
    Occupational Scales, 175  
    Report, 175  
    scales for, 174  
Cancellation subtest, 111, 117–118  
CAPA. *See* Child and Adolescent  
    Psychiatric Instrument (CAPA)  
CAPI V21.1.3. *See* Computer Assisted  
    Personal Interview V21.1.3 (CAPI  
    V21.1.3)  
Care for patients, 579–581  
Career Ability Placement Survey (CAPS),  
    147  
Career Assessment Inventory (CAI), 170,  
    173, 180–181  
Career exploration, 184  
Career Occupational Preference System  
    Interest Inventory (COPS), 147

- Career Orientation Placement and Evaluation Survey (COPES), 147
- Caregiver Assessment of Function and Upset (CAFU), 507–508
- Caregivers, 449–450
- Carl Hollow Square test, 243
- CASI. *See* Child Anxiety Sensitivity Index (CASI); *Children's Attributional Style Interview* (CASI)
- CAT. *See* Computer adaptive testing (CAT)
- Category Test, 242–243, 250
- Cattell–Horn–Carroll theory (CHC theory), 32–33, 75, 82–83, 92–93, 205–206
- second-order factors and third-order clusters, 208
- Causal
- directionality, 467t
  - mediation, 467t
  - presuppositions, 490
  - relations, 467t, 490–491
  - variables, 489–490
- CBC-DOF. *See* Child Behavior Checklist Direct Observation Form (CBC-DOF)
- CBC-TRF. *See* Child Behavior Checklist Teacher Report Form (CBC-TRF)
- CBCL. *See* Child Behavior Checklist (CBCL)
- CCAB. *See* Complex Cognitive Assessment Battery (CCAB)
- CD-RISC. *See* Connor–Davidson Resilience Scale (CD-RISC)
- CDR. *See* Clinical Dementia Rating (CDR)
- CDSS. *See* Calgary Depression Scale for Schizophrenia (CDSS)
- CELF-5. *See* Clinical Evaluation of Language Fundamentals—Fifth Edition (CELF-5)
- Cerebral concussion, 288
- Change Sensitive Scores (CSS), 54
- CHC theory. *See* Cattell–Horn–Carroll theory (CHC theory)
- Checklists, 169–170, 451–452
- Child and Adolescent Psychiatric Instrument* (CAPA), 342–343
- Child Anxiety Sensitivity Index (CASI), 447
- Child Attention-Deficit Hyperactivity Disorder Teacher Telephone Interview* (CHATTI), 347
- Child Behavior Checklist (CBCL), 443–444
- Child Behavior Checklist Direct Observation Form (CBC-DOF), 441–442
- Child Behavior Checklist Teacher Report Form (CBC-TRF), 443–444
- Child PTSD Symptom Scale for DSM-5* (CPSS-5), 346
- Child(ren)
- behavior, 200–202
  - covert processes assessment, 446–449
  - direct measures, 446–447
  - interviews, 446
  - self-monitoring, 448–449
  - self-report instruments, 447–448
- depression, 347
- historical perspectives of interviews, 337–339
- neuropsychological evaluations, 193
- historical foundation, 194–198
  - origins, 195–198
  - process, 198–209
  - psychoeducational vs., 199–201
  - sources of error, 209–219
- overt behavior assessment, 441–445
- behavior ratings and checklists, 443–444
  - direct behavioral observation, 441–443
  - interviews, 444–445
  - structured interviews, 339–348
  - strengths and limitations, 348–349
- Children's and Adult Color Trails Test, 556
- Children's Attributional Style Interview* (CASI), 347
- Children's Interview for Psychiatric Syndromes* (ChIPS), 343
- Children's Report of Parenting Behavior Inventory (CRPBI), 452
- CHS. *See* Clinical History Schedule (CHS)
- CICS. *See* Coping Inventory for Competitive Sport (CICS)
- CIDI. *See* Composite International Diagnostic Interview (CIDI)
- CIRVs. *See* Composite International Reference Values (CIRVs)
- CISG. *See* Concussion in Sport Group (CISG)
- CISS. *See* Campbell Interest and Skill Survey (CISS); Coping Inventory for Stressful Situation (CISS)

- Clandestine behaviors, 482  
Classical conditioning, 438  
Classical Test Theory (CTT), 39, 44–45  
Classroom Observation Code, 442  
Client variance, 317  
Client-centered approach, 310–314  
Clinical assessment strategies, 203–204, 465  
  integrating multiple measures, 488–489  
  psychometric principles, 470*t*  
  self-monitoring, 484  
Clinical decision making, 216–219, 217*t*  
Clinical Dementia Rating (CDR), 514–515  
Clinical Evaluation of Language  
  Fundamentals—Fifth Edition  
  (CELF-5), 446–447  
Clinical History Schedule (CHS), 367–368  
Clinical interview, 307–316, 311*t*, 337, 355.  
  *See also* Diagnostic interviews  
  culture and diversity, 323–325  
  DSM-5, 328–330  
  format selection  
    flexibly structured interviews, 319–321  
    free-format or open interviews,  
      318–319  
    reliability and validity, 317–318  
    structured and semistructured  
      interviews, 321–323  
  structure, 315–316  
  technology, 325–327  
Clinical Neuropsychology, 276–277  
Clinical practice, achievement testing in, 151  
Clinical reasoning, 327  
Clinical science in psychological assessment, 461–496  
CLNT. *See* Cross-Linguistic Naming Test (CLNT)  
Cluster analysis, 173, 240  
CNEHRB system. *See* Comprehensive  
  Norms for an Extended  
  Halstead–Reitan Battery system  
  (CNEHRB system)  
CO subtest. *See* Comprehension subtest (CO  
  subtest)  
Coding subtest, 117  
Cognition, 554, 580–581  
Cognitive abilities, 505–506  
Cognitive behaviorists, 438–439  
Cognitive domains, 581–582  
Cognitive functioning  
  dementia, 512–513  
  assessment instruments, 514–515  
  establishment of premorbid functioning,  
    514  
Cognitive processes, 73–74, 554–555  
  involving in reading achievement,  
    158–159  
Cognitive Proficiency Index (CPI), 118–119  
Cognitive rehabilitation, 583–585  
Cognitive social learning person variables  
  (CSLPVs), 439  
Cognitive training, computers for, 584  
Cognitive variability, 134–135  
Cognitive/verbal responses, 472–473  
Collaborative & Proactive Solutions model  
  (CPS model), 449–450  
Collateral contacts, 537  
Collateral sources of information, 539  
College, aptitude and achievement testing in,  
  144–146  
Communication, 559–561  
Competencies, 439  
Complex Cognitive Assessment Battery  
  (CCAB), 577  
Complexity, 427–428  
Composite International Diagnostic  
  Interview (CIDI), 324–325, 356, 360  
Composite International Reference Values  
  (CIRVs), 425–426  
Comprehension subtest (CO subtest), 85,  
  111–113  
Comprehensive assessment, 16, 122, 200,  
  307–308  
Comprehensive batteries, 241–262  
  HRB, 242–244  
  Luria–Nebraska Neuropsychological  
    Battery, 255–262  
Comprehensive cognitive assessments  
  ACS Social Cognition, 129–130  
  ACS Suboptimal Effort, 130  
  Binet–Simon Scale, 104  
  expanded assessment, 122  
  five-factor models, 120–121  
  future directions, 135  
  issues in summarizing overall ability,  
    119–120  
  joint factor structure of WAIS-IV and  
    WMS-IV, 128–129  
  subtest level changes, 108–111

- WAIS-IV, 106  
development approach, 106–108  
and digital assessment, 121–122  
index scores and structure, 111–119  
refining interpretation, 130–135
- WMS-IV, 123–128
- Comprehensive neuropsychological test battery, 227–228
- Comprehensive Norms for an Extended Halstead–Reitan Battery system (CNEHRB system), 244
- Comprehensive System (CS), 423–426
- Comprehensive Test of Phonological Processing (CTOPP), 158
- Comprehensiveness, 435–436
- Computational assessment/neuropsychology, 582–583
- Computer adaptive testing (CAT), 45
- Computer administered clinical interviews and ratings, 327
- Computer Assisted Personal Interview V21.1.3 (CAPI V21.1.3), 360
- Computer-automated assessment of multitasking, 577
- Computerized tomographic scanning (CT scanning), 581–582
- Computers for cognitive training, 584  
ecological momentary assessments, 585  
smartphones for psychological assessment, 584–585
- Concussion assessment, 288–289  
off-field evaluations, 290–292  
sideline evaluations, 290–295
- Concussion in Sport Group (CISG), 288
- Conditional behavior, 465
- Conditioning, 438
- Confirmation bias, 71
- Connor–Davidson Resilience Scale (CD-RISC), 285
- Construct validation of neuropsychological tests, 240
- Constructional abilities, 113, 518
- Contemporaneous causal relations, 465, 467t
- Contemporary intelligence tests, 207–209
- Contemporary self-report inventories, 398
- Content Grouping with Statistical Refinement Method, 399t, 401, 408–409
- Content-based item selection method, 399t, 404–406
- Content-sampling error, 55, 57
- Context, 435–436, 440  
assessment, 449–453  
checklists, 451–452  
cultural considerations, 452–453  
interviews, 449–451  
of testing, 552
- Contextual behavior, 465, 466t
- Contextual issues in forensic psychology assessment, 538
- Contingency fees, 543
- Control of variance, 204
- COPES. *See* Career Orientation Placement and Evaluation Survey (COPES)
- Coping Inventory for Competitive Sport (CICS), 285
- Coping Inventory for Stressful Situation (CISS), 284–285
- Coping skills, 284–285
- COPS. *See* Career Occupational Preference System Interest Inventory (COPS)
- COPSSystem, 147
- Courtroom testimony, 541–542, 541t
- Covert behaviors, 435, 440
- CPI. *See* California Psychological Inventory (CPI); Cognitive Proficiency Index (CPI)
- CPS model. *See* Collaborative & Proactive Solutions model (CPS model)
- CPSS-5. *See* Child PTSD Symptom Scale for DSM-5 (CPSS-5)
- Creating (Orientation Scale), 174
- Credibility, 278
- Criterion keying method, 399–400
- Criterion variance, 317
- Criterion-referenced score interpretations, 53
- Criticisms, 77–79, 160–161, 318–319
- Cross-battery assessment methods, 207–209
- Cross-cultural comparison, 555–556
- Cross-cutting symptom measures, 328–329
- Cross-Linguistic Naming Test (CLNT), 556
- Cross-sectional methods, 182, 506
- CRPBI. *See* Children's Report of Parenting Behavior Inventory (CRPBI)
- CS. *See* Comprehensive System (CS)
- CSLPVs. *See* Cognitive social learning person variables (CSLPVs)

- CSS. *See* Change Sensitive Scores (CSS)
- CT scanning. *See* Computerized tomographic scanning (CT scanning)
- CTOPP. *See* Comprehensive Test of Phonological Processing (CTOPP)
- CTS2. *See* Revised Conflict Tactics Scale (CTS2)
- CTT. *See* Classical Test Theory (CTT)
- Cultural considerations, 452–453
- Cultural Formulation, 554
- Culture, 211–213, 465–471, 558  
and diversity, 323–325  
proxies, 554
- “Culture-fair” tests, 555–556
- “Culture-free” tests, 556
- Curiosity, 197–198
- Cybernetic processing, 66
- Cynicism scale (CYN), 409<sup>t</sup>
- D**
- DAT. *See* Differential Aptitude Test (DAT)
- Data analysis, intuitive, 481
- Daubert test, 537
- Daubert v. Merrell Doe Pharmaceuticals standard, 536
- De facto* tests, 555
- Decision-making process, 103, 488  
expectable biases in clinical, 217<sup>t</sup>  
personality assessment procedures for, 11
- Declarative memory, 507
- Delayed Memory, 127
- Delayed Memory Index (DMI), 127
- Delis–Kaplan Executive Functioning System, 80–81
- Delusions, 320, 375–377, 515, 519, 538
- Dementia, 7–8, 510, 512–513, 520–521, 521<sup>t</sup>  
assessment instruments, 514–515  
family history, 509  
syndrome of depression, 512
- Dementia Rating Scale (DRS), 514–515
- Demographic referenced norms, 131
- Deposition, 541, 546
- Depression (*D*), 369–370, 406, 407<sup>t</sup>, 505, 510–512, 520–521, 521<sup>t</sup>  
behavior, 340–341  
depression-related cognitive dysfunction, 512  
disorder, 17
- Geriatric Depression Scale, 511<sup>t</sup>  
scale, 409<sup>t</sup>
- Developmental and cultural norms, 435
- Diagnostic and Statistical Manual of American Psychiatric Association, 18–19
- Diagnostic and Statistical Manual of Mental Disorders (DSM), 18–19, 323–324, 355, 512, 538–539
- DSM-III, 321, 338–339
- DSM-5, 328–330, 513  
criteria for major neurocognitive disorder, 513<sup>t</sup>
- Diagnostic assessments, 435–436
- Diagnostic impressions, 538–539
- Diagnostic Interview for Children and Adolescents* (DICA), 341–342
- Diagnostic Interview for Children-version IV* (DISC-IV), 341, 358–359
- Diagnostic Interview Schedule (DIS), 7, 321, 355–356, 358–360
- Diagnostic Interview Schedule for Children (DISC), 359–360
- DISC-IV, 445
- Diagnostic interviews, 338–339, 356–362.  
*See also* Clinical interview;  
Structured interviews
- CIDI, 360
- DIS, 358–360
- IPDE, 360–361
- MINI, 358
- SCID-5-PD, 361–362
- SCID-5, 357–358
- Diagnostic Psychological Testing*, 421
- DICA. *See* *Diagnostic Interview for Children and Adolescents* (DICA)
- Dichotic listening, 230–231
- Dictionary of Occupational Titles (DOT), 181
- DIF. *See* Differential item function (DIF)
- Differential Aptitude Test (DAT), 147
- Differential item function (DIF), 45–46
- Difficulty handling complex tasks, 513
- Difficulty learning new information, 513
- Digit Span, 115
- Digit Span Backward (DSB), 115
- Digit Span Forward (DSF), 115
- Digit Span Sequencing (DSS), 115
- Digit Vigilance Test, 246

- Digit-Symbol Coding on WAIS-III, 117  
Digital assessment, 121–122  
Direct behavioral observation, 472  
Direct testimony, 541–542, 542<sub>t</sub>  
**DIS.** *See* Diagnostic Interview Schedule (DIS)  
Disabilities Education Improvement Act (IDEA), 157  
**DISC.** *See* Diagnostic Interview Schedule for Children (DISC)  
**DISC-IV.** *See* *Diagnostic Interview for Children-version IV* (DISC-IV)  
Disconfirmatory approach, 199  
Discriminative validity studies, 259  
Disorganization, 366, 375–376  
Disruptive behavior disorders, 345–346  
Dissimulation scales, 36–37  
Distractibility, 279  
Distress domain, 379  
Diversity panel, 42  
**DMI.** *See* Delayed Memory Index (DMI)  
Documentation in forensic psychology, 545–546  
Documents reviewing in forensic psychology assessment, 537  
**DOT.** *See* Dictionary of Occupational Titles (DOT)  
Double discrimination scales, 256–257  
Double dissociation, 233–234  
**DPICS.** *See* Dyadic Parent–Child Interaction Coding System IV (DPICS)  
**DRS.** *See* Dementia Rating Scale (DRS)  
**DSB.** *See* Digit Span Backward (DSB)  
**DSF.** *See* Digit Span Forward (DSF)  
**DSM.** *See* Diagnostic and Statistical Manual of Mental Disorders (DSM)  
**DSS.** *See* Digit Span Sequencing (DSS)  
Duration recording, 477  
Dyadic Parent–Child Interaction Coding System IV (DPICS), 442  
Dynamic behavior, 465, 466<sub>t</sub>  
Dynamic causal relations, 465, 467<sub>t</sub>  
Dynamic Visual Field Test, 243  
Dysphoria factor, 370, 374–375, 511–512  
Dysphoric manic episodes, 370
- E**  
“E-health”, 325–326  
Early Development and Home Background Form—Clinician (EDHB-C), 330
- Early Development and Home Background Form—Parent/Guardian (EDHB-PG), 330  
**EBM.** *See* Evidence-based medicine (EBM)  
Ecological momentary assessment (EMA), 325–326, 464, 484–485, 585  
Ecological validity, 213, 480  
Education(al), 562  
    alternatives, 177  
    education-based norms, 562–563  
    quality, 563  
    recommendations, 200  
*Educational Opportunities Finder, The*, 177  
Educational quality measure (QEd), 155  
**EEG.** *See* Electroencephalogram (EEG)  
**Ego Impairment Index (EII)**, 428  
    EII-3, 428  
“Elated mania”, 374–375  
Electroencephalogram (EEG), 424  
**EMA.** *See* Ecological momentary assessment (EMA)  
Embedded measures, 215–216  
**EMBRACED.** *See* European domain-specific computerized battery for cross-cultural comparisons (EMBRACED)  
Emotion regulation, 283–284  
Emotional expressivity (EXP), 379–380  
Empirically derived method, 399–400, 399<sub>t</sub>  
Employment, 146–147  
    testing, 144  
Enactment analogs, 482–483  
Encoding strategies, 439  
English-dominant bilinguals, 560  
Environmental events, 485  
    factor, 286–287  
Epidemiologic Catchment Area Program, 359–360  
Error variance, 204, 474–475  
Esquirol’s development of methods, 103  
**ESSA.** *See* Every Student Succeeds Act (ESSA)  
Ethical issues in forensic psychology, 543–544  
Ethnicity, 211, 554  
Euphoric mania, 370  
“Euphoric-grandiose” factor, 374–375

- European domain-specific computerized battery for cross-cultural comparisons (EMBRACED), 556
- Event sampling, 476–477
- Every Student Succeeds Act (ESSA), 150
- Evidence-based intervention strategies, 203
- Evidence-based medicine (EBM), 70–71
- Executive functioning, 517
- Exercise science, 277
- Exner, John, 423
- EXP. *See* Emotional expressivity (EXP)
- “Expanded Halstead–Reitan Battery”, 246
- Expectancies, 439
- Experimental manipulation, 491
- Expert witness, 536  
fees, 543
- Expertise abilities, 206, 207<sup>t</sup>
- Explicit memory. *See* Declarative memory
- Exploratory factor analysis. *See* Factor-analytic strategy
- Exploratory process, 436–437
- Expressive Speech scales, 256, 258
- Extended social systems, 465, 467<sup>t</sup>
- Extensive factor-analytic studies, 257
- Extinction, 230
- Extraversion scale measures, 175
- F**
- F-scales, 36–37
- FACCD. *See* Functional Analytic Clinical Case Diagram (FACCD)
- Face validity, 171–172, 239
- Face-to-face verbal exchanges, 308
- Factor analysis, 5, 173, 242, 369–370, 427, 511–512
- Factor scales, 256–257
- Factor scores, 209–210, 366
- Factor-analytic strategy, 239–240, 399<sup>t</sup>
- Fair treatment, 552
- Fairness in psychological testing, 551–553  
importance in testing minority individuals, 553–555  
variables contributing to test performance  
in minorities, 555–563  
acculturation and assimilation, 558–559  
communication and language, 559–561  
psychological construct, 555–557  
SES, 562–563
- test norms, 557–558
- Fake Bad Scale (FBS), 404–406
- “Fake Good” scales, 37
- Falsification, 199
- Familiarity, 552
- Family Adaptability and Cohesion Evaluation Scales-II, 451
- Family Environment Scale, 451
- Family problems scale (FAM), 409<sup>t</sup>
- FBA. *See* Functional behavioral assessment (FBA)
- Fear Survey Schedule for Children-Revised (FSSR-R), 447–448
- Fears scale (FRS), 409<sup>t</sup>
- Figure Memory Test, 246
- Figure Weights (FW), 110, 113–114
- Financial Services Basic Scales, 174
- Fine Arts-Mechanical scale, 181
- Finger tapping, 229, 243, 245, 249
- First-order human-cognitive abilities, 206
- 5-point Likert scale, 488
- “Fixed battery” approach, 219, 243
- Flexibility, 275
- Flexible battery approach, 219
- Flexibly structured interviews, 319–321
- Flicker fusion  
procedure, 243  
tests, 248–249
- Flight Attendant, 171
- Fluency, reading disorders, 159
- Fluid process, 436–437
- Fluid Reasoning Index (FRI), 85, 107–108
- Fluid reasoning subdomain, 113
- Flynn effect, 32
- fMRI technology. *See* Functional magnetic resonance technology (fMRI technology)
- Forced-choice measures, 215
- Forensic psychology, 533–534. *See also* Neuropsychology; Psychology  
assessment tools, 539–540  
collateral sources of information, 539  
courtroom testimony, 541–542, 541<sup>t</sup>  
dealing with subpoenas, 546  
documentation and record keeping, 545–546  
ethical issues, 543–544  
expert witness fees, 543  
forensic process, 537–539

- forensic roles, 535–536, 540–541  
legal courts system, 534–535  
multiple relationships, 544  
U.S. legal decisions on scientific expertise, 536–537  
working within legal system, 544–545
- Form quality research, 425–426
- Formal cognitive assessment, 556–557
- Formal interviews, 445
- FQ%. *See* Percentage of minus form quality answers (FQ%)
- Free-format interview, 318–319
- Freedom from Distractibility Index, 107–108
- Frequency, 426
- of Behavior Problems Scale, 507–508
- FRI. *See* Fluid Reasoning Index (FRI)
- FRS. *See* Fears scale (FRS)
- Frye test, 537
- FSSR-R. *See* Fear Survey Schedule for Children-Revised (FSSR-R)
- Full Scale IQ (FSIQ), 84–85, 119–120
- Fully structured interviews, 321–322
- Functional analysis
- constructing functional analysis and FACCD, 492–493
  - integrating multiple measures in clinical assessment into, 488–489
  - interventions effects based or not based on, 496
- Functional Analytic Clinical Case Diagram (FACCD), 488–489, 492–493
- Functional behavioral assessment (FBA), 496
- Functional behavioral interviews and questionnaires, 485–486
- Functional Diagnostic Protocol*, 347
- Functional magnetic resonance technology (fMRI technology), 234–235, 423–424
- FW. *See* Figure Weights (FW)
- G**
- Ga* factor, 92
- GAI. *See* General Ability Index (GAI)
- GATB. *See* General Aptitude Test Battery (GATB)
- “Gatekeeper” operation, 203
- Gaussian curve, 52
- Gc* factor, 89
- GDS. *See* Geriatric Depression Scale (GDS)
- General Ability Index (GAI), 118–119
- General Aptitude Test Battery (GATB), 147
- General Intellectual Ability (GIA), 87–89
- General Interest Survey, 170
- General Neuropsychological Deficit Scale (GNDS), 244–247
- General Occupational Themes (GOT), 173, 178
- General Reference Sample, 175
- General Themes and Basic Interest Areas, 180–181
- Generalizability, 204
- Geriatric assessments, special problems in, 522
- Geriatric Depression Scale (GDS), 511–512, 511t
- Geropsychology, 505
- Gerstmann Syndrome, 252–253
- Gf* factor, 89–92
- GIA. *See* General Intellectual Ability (GIA)
- Glasgow Coma Scale, 237–238
- Glr* factor, 92
- GNDS. *See* General Neuropsychological Deficit Scale (GNDS)
- GOT. *See* General Occupational Themes (GOT)
- Grade equivalents, 53
- Grade point average (GPA), 58, 144
- Grasping reflex, 197
- Gross motor behavior, 442
- Group interviews, 451
- Grundzüge der Physiologischen Psychologie*, 276
- Gs* factor, 92
- Gv* factor, 92
- Gwm* factor, 92
- H**
- Hallucinations, 375–377
- Halstead Category Test (HCT), 245, 576–577
- Halstead Impairment Index, 256–257
- Halstead Tactual Performance Test, 245
- Halstead’s biological intelligence tests, 245
- Halstead–Reitan battery (HRB), 234, 241–244, 574. *See also* Luria–Nebraska Neuropsychological Battery (LNNB)

- Halstead–Reitan battery (HRB) (*Continued*)  
evaluation, 251–254  
history, 242–244  
standardization research, 248–251  
structure and content, 244–254  
theoretical foundations, 247–248
- Halstead–Russell Neuropsychological Evaluation System (HRNES), 244
- Hamilton Rating Scale for Depression (HRSD), 371–372
- HCT. *See* Halstead Category Test (HCT)
- Health concerns scale (HEA), 409*t*
- Healthcare delivery system, 325–326
- Healthy responses of MOA (MAH), 429
- Helping (Orientation Scale), 174
- Henmon–Nelson tests of mental ability, 243
- Heterochrony, 196–197
- Heterogeneous behavior, 466*t*
- Heterogeneous scales, 172–173  
current interest inventories, 173–174  
development, 173–174
- High-Level Language Aptitude Battery, 148–149
- Home Situations Questionnaire (HSQ), 450
- Home Situations Questionnaire and revision (HSQ-R), 450
- Homogeneous scales, 172–173  
development, 173
- Hooper Visual Organization Test, 518
- HRB. *See* Halstead–Reitan battery (HRB)
- HRNES. *See* Halstead–Russell Neuropsychological Evaluation System (HRNES)
- HRSD. *See* Hamilton Rating Scale for Depression (HRSD)
- HSQ. *See* Home Situations Questionnaire (HSQ)
- HSQ-R. *See* Home Situations Questionnaire and revision (HSQ-R)
- Human cognitive functioning, 206–207
- Human intelligence, 65–66
- Human neuropsychology laboratory, 247
- Huntington’s disease, gene for, 236–237
- ‘Hybrid’ neuropsychological testing approach, 292
- Hypochondriasis Scale (*Hs* Scale), 406, 407*t*
- Hypothesis-testing process, 436–437
- Hysteria (*Hy*), 32–33, 406, 407*t*
- I**
- ICC. *See* Item characteristic curves (ICC)
- ICD. *See* International Classification of Diseases (ICD)
- “Iceberg” profile, 283, 283*f*
- ICI-M. *See* Interactive Computer Interview for Mania (ICI-M)
- IDEA. *See* Disabilities Education Improvement Act (IDEA)
- Ideational impairment, 428
- Idiographic  
approach, 472  
assessment strategies, 464
- IDS. *See* Inventory of Depressive Symptomatology (IDS)
- IGC. *See* Item Group Checklist (IGC)
- IIP. *See* Interest Item Pool (IIP)
- Illegal behaviors, 482
- Immediate Memory Index (IMI), 126–127
- Immediate Post-concussion Assessment and Cognitive Testing (ImPACT), 293, 580–581
- Impaired reasoning, 513
- Impaired spatial ability and disorientation, 513
- Impairment index, 242, 244–247
- Implicit memory. *See* Nondeclarative memory
- Inconsistency scales, 37
- Incremental validity, 209–211
- Independent Medical Examination (IME), 535
- Index of Teaching Stress (ITS), 452
- Index(es), 209–210  
scores, 106–107
- Individual achievement test-revised–normative update (PIAT-R/NU), 157–158
- Individual Zone of Optimal Function (IZOF), 284
- Individualized approach to interpretation of test scores, 553
- Inflection point, 45
- Influencing (Orientation Scale), 174
- Informal interviews, 445
- Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE), 507–508
- Informant report, 317

- Information (IN)  
subtest, 111–112  
variance, 317
- Informed consent, 492, 533  
for forensic psychology assessment, 537
- “Infrequency” scales. *See* F-scales
- “Inhibition”, 473
- Inkblots, 419, 422
- Insomnia, 363–364
- Instrument assessment, 470*t*
- Intellectual  
assessment, 103–104  
disabilities, 103  
Processes scales, 256
- Intelligence, 65–67, 82–83  
and achievement testing, 4–5  
framework for interpreting intelligence tests, 70–71  
individually administered tests, 82  
tests, 5, 67–69, 82
- Intelligence quotient (IQ), 5, 50–51, 68, 104, 205, 446–447
- Intelligere*, 67
- Interactive causal paths, 465, 467*t*
- Interactive Computer Interview for Mania (ICI-M), 327
- Interest Analysis Blank, 170
- Interest inventories, 169–170  
CAI, 180–181  
career exploration, 184  
characteristics of good items, 170–172  
CISS, 174–176  
construction, 172–173  
earliest item pool, 170  
future directions, 185–186  
Holland’s, 176–177  
influence of vocational interest theories, 172–173
- O\*NET Interest Profiler and Interest Item Pool, 181–183  
reliability and validity, 177  
research, 184–186  
selection and placement, 184  
SII, 177–179  
use of, 183
- Interest Item Pool (IIP), 173–174, 181–183
- Interest Profiler, The, 181
- Interest Report Blank, 170
- Interference, 442
- Interfunctional development, 197
- Internal consistency coefficients, 56–57
- Internal structure specification, 36–37
- International Activities, 174
- International Classification of Diseases (ICD), 538–539  
ICD-10, 358
- International Personality Disorder Examination (IPDE), 356, 360–361
- International Personality Item Pool (IPIP), 173–174
- Interpersonal  
factors, 307–308  
variables, 489
- Interpreters for testing, 561
- Interrater  
agreement, 57  
or interscorer differences, 55  
reliability, 57–58, 344
- Interscorer reliability, 57
- Interval recording, 477
- Interval scales, 49–50
- Interventions, 203
- Interview, 6–8, 310, 342, 444–446, 449–451. *See also* Clinical interview; Diagnostic interviews; Structured interviews  
free-format, 318–319  
open, 318–319  
semistructured, 321–323  
style, 307
- Interviewers, 338
- Interviewing, 539
- Intrafunctional development, 197
- Inventory of Depressive Symptomatology (IDS), 371–373
- IPDE. *See* International Personality Disorder Examination (IPDE)
- IPIP. *See* International Personality Item Pool (IPIP)
- IQ. *See* Intelligence quotient (IQ)
- IQCODE. *See* Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE)
- “Irritable mania”, 374–375
- IRT. *See* Item Response Theory (IRT)
- Item characteristic curves (ICC), 44–45, 44*f*, 45*f*, 46*f*
- Item difficulty index, 43

- Item discrimination, 43–44  
 Item Group Checklist (IGC), 367–368  
 Item pool and profile  
   for CISS, 174  
   of SII, 178  
 Item Response Theory (IRT), 42–46, 54, 108–109  
*ITS. See Index of Teaching Stress (ITS)*  
 IZOF. *See Individual Zone of Optimal Function (IZOF)*
- J**  
 Job  
   design, 185–186  
   interviews, 308  
 Joint factor structure of WAIS-IV and WMS-IV, 128–129  
 “Junk science” testimony, 536
- K**  
*K-SADS-PL. See Schedule for Affective Disorders And Schizophrenia For School-Aged Children Present And Lifetime Version (K-SADS-PL)*  
 KeyMath-3 Diagnostic Assessment, 162  
*KID-SCID. See Structured Clinical Interview for DSM-IV Childhood Diagnoses (KID-SCID)*  
 Klove Grooved Pegboard Test, 243, 246  
 Klove Roughness Discrimination Test, 246  
 Knowledge, 37–39, 112, 115, 147, 204–205  
 Kohs Blocks, 230  
 Kuder–Richardson Formula 20 (KR 20), 56–57
- L**  
*L-scales. See “Lie Scales” (L-scales)*  
 Language, 211–213, 554, 559–561  
   acquisition, 148–149  
   dominance index, 560  
   forms of language assessment, 240  
   impairment, 513  
   neuropsychological assessment, 231  
   proficiency, 560  
   recognition, 582–583  
 Leadership scale, 179  
 Learning, 516  
   disorders, 83–84, 143–144
- Environment scale, 179  
 Learning disability (LD), 75–76, 118  
 Left Hemisphere scales, 256  
 Legacy  
   instruments, 339–340  
   semistructured interview, 341  
 Legal courts system, 534–535  
   working within, 544–545  
*Leisure Activity Finder, The*, 177  
 Leisure Interest Questionnaire (LIQ), 173–174, 185  
 Letter–Number Sequencing  
   subtest, 116  
   trials, 109  
 Letter–pattern matching, 92  
 “Lie Scales” (L-scales), 37  
 Life-saving interventions, 202  
 Litigation, 404–406  
*Litigation Response Syndrome (LRS)*, 404–406  
 LNNB. *See Luria–Nebraska Neuropsychological Battery (LNNB)*  
 Localization scales, 256–257  
 Logical Memory, 125  
   subtest, 507  
 Longitudinal  
   methods, 182  
   studies, 237–238  
 Loss of speed in performing skilled activities, 229  
 Low self-esteem scale (LSE), 409*t*  
 Lowest “Like” response rate, 170, 171*f*  
*LRS. See Litigation Response Syndrome (LRS)*  
*Luria–South Dakota Neuropsychological Battery*, 255  
*Luria’s Neuropsychological Investigation*, 255  
*Luria’s tests*, 258–259  
*Luria–Nebraska Battery*, 240–241  
*Luria–Nebraska Neuropsychological Battery (LNNB)*, 234, 255–262. *See also Halstead–Reitan battery (HRB)*  
   evaluation, 260–261  
   history, 255  
   standardization research, 259–260  
   structure and content, 256–257  
   theoretical foundations, 257–259

**M**

- m-BESS. *See* Modified Balance Error Scoring System (m-BESS)
- Ma.* *See* Mania (*Ma*)
- MAC. *See* Memory Assessment Clinics (MAC)
- MacAndrew alcoholism scale (MAC-R), 408<sup>t</sup>
- MADRS. *See* Montgomery Asberg Depression Rating Scale (MADRS)
- Magnetic resonance imaging (MRI), 234–235, 581–582
- Magnetic resonance spectroscopy (MRS), 234–235
- MAH. *See* Healthy responses of MOA (MAH)
- Mail-in-scoring service, 174
- Main-Stream Code for Instructional Structure and Student Academic Response (MS-CISSAR), 442
- Major neurocognitive disorder, 513
- DSM-5 criteria, 513<sup>t</sup>
- Major personality assessment method (MMPI), 246, 250
- Major psychological assessment paradigms, 463
- Majority group, 551
- Make-a-picture-story test (MAPS test), 8–9
- Making a Speech*, 170
- Malingering, 214–216
- Mania (*Ma*), 406, 407<sup>t</sup>
- MAP. *See* Motivation and pleasure (MAP); Pathological responses of MOA (MAP)
- Marital distress scale (MDS), 408<sup>t</sup>
- Marker variable strategy, 491
- Martens' model of personality structure, 282<sup>f</sup>
- MAS. *See* Bech–Rafaelsen Mania Scale (MAS)
- Masculinity/Femininity (*Mf*), 406, 407<sup>t</sup>
- Mastery testing, 53
- Mathematical achievement, current measures of, 161–162
- Mathematics and Science Basic Scales, 174
- Matrix Reasoning (MR), 109, 113–114
- Maximization of systemic variance, 204
- MCAT. *See* Medical College Admission Test (MCAT)

- MCI. *See* Mild Cognitive Impairment (MCI)
- MCMI. *See* Millon Clinical Multiaxial Inventories (MCMI)
- MDS. *See* Marital distress scale (MDS)
- Meaning of neuropsychological evaluation, 201–209
- Measured construct, access to, 553
- Measurement
- bias, 552
  - error, 470<sup>t</sup>
- Mediators, 465, 467<sup>t</sup>
- Medical
- conditions, 508–509
  - history, 508
- Medical College Admission Test (MCAT), 145–146
- Medicare, 545
- Medication history, 508
- Memory, 239, 516
- assessments, 507
  - impairment, 228–229
  - scales, 256, 260–261
  - testing, 240
- Memory Assessment Clinics (MAC), 586
- Mental
- capability, 65–66
  - processing, 66
  - retardation, 5
  - status, 310–314, 320, 436–437
  - tests, 3–4, 397
  - toughness, 280–281, 286
  - higher mental functions, 195
- Mental status examination, 436–437
- Mental Toughness Scale (MTS), 286
- MEPS. *See* Military Entrance Processing Stations (MEPS)
- MET. *See* Military Entrance Test site (MET)
- Methodological issues, 506
- Meyers Neuropsychological System, 241
- Meyers procedure, 242
- Mf.* *See* Masculinity/Femininity (*Mf*)
- Michigan Battery, 241
- Mild Cognitive Impairment (MCI), 237–238
- Mild traumatic brain injury (MTBI), 288
- Milieu variables, 489
- Military Entrance Processing Stations (MEPS), 148
- Military Entrance Test site (MET), 148

- Military Occupational Classifications (MOC), 177
- Military testing for identification and classification, 148
- Military/Law Enforcement, 174
- Millon Adolescent Personality Inventory tests, 10–11
- Millon Behavioral Health Inventory tests, 10–11
- Millon Clinical Multiaxial Inventories (MCMI), 401
- PsycINFO publications on, 410–412, 410<sub>t</sub> tests, 10–11
- Mindfulness-based behavior therapy, 495–496
- Mini Interest Profiler (Mini-IP), 181–182
- Mini International Neuropsychiatric Interview (MINI), 324–325, 356, 358
- Mini International Neuropsychiatric Interview for Children and Adolescents* (MINI-KID), 343–344
- Mini Mental Status Examination (MMSE), 359, 514–515, 518, 522
- Minnesota Interest Inventory, 170
- Minnesota Multiphasic Personality Inventory (MMPI), 9–10, 20–21, 32–33, 282–283, 400–401
- clinical scales, 399–400
- MMPI-2-RF, 402–403, 411–412
- PsycINFO publications on, 410–412, 410<sub>t</sub>
- MMPI-2, 10–11
- assessing protocol validity with, 404–406
- clinical scales, 402–403, 406, 407<sub>t</sub>
- content scales, 408–409, 409<sub>t</sub>
- PsycINFO publications on, 410–412, 410<sub>t</sub>
- RC Scales, 399<sub>t</sub>, 401–402
- scale, 400–401, 403–406
- validity scales, 405<sub>t</sub>
- supplementary scales, 407–408, 408<sub>t</sub>
- Minor motor movement, 442
- Minority, 553
- groups, 553
- Mirror neurons, 424
- MMPI. *See* Major personality assessment method (MMPI)
- MMSE. *See* Mini Mental Status Examination (MMSE)
- MOA. *See* Mutuality of Autonomy (MOA)
- Mobile electronic devices, 484
- MOC. *See* Military Occupational Classifications (MOC)
- Moderator variables, 465, 467<sub>t</sub>
- Modern clinical child interviews, 339
- Modification, 555–556
- Modified Balance Error Scoring System (m-BESS), 290–291
- Modified story memory technique, 584
- Monocultural instruments, 554
- Monolingual instruments, 554
- Montgomery Asberg Depression Rating Scale (MADRS), 372
- Montreal Cognitive Assessment, 7–8
- Mood disorders, 8
- Motivation and pleasure (MAP), 376–377
- Motivation Assessment Scale, 485
- Motivation(al)
- analysis, 16
- interviewing, 8, 314
- Motor
- scales, 256
- skills, 517
- MR. *See* Matrix Reasoning (MR)
- MRI. *See* Magnetic resonance imaging (MRI)
- MRS. *See* Magnetic resonance spectroscopy (MRS)
- MS. *See* Multiple sclerosis (MS)
- MS-CISSAR. *See* Main-Stream Code for Instructional Structure and Student Academic Response (MS-CISSAR)
- MTBI. *See* Mild traumatic brain injury (MTBI)
- MTS. *See* Mental Toughness Scale (MTS)
- Multidimensional inventories of psychopathology
- assessing protocol validity with MMPI-2, 404–406
- assessing psychopathology
- MMPI-2 clinical scales, 406
- MMPI-2 content scales, 408–409
- MMPI-2 supplementary scales, 407–408
- development and uses, 399–404

- precursors to self-report inventories, 397–398
- publications on MMPI-2, MMPI-2-RF, PAI, and MCMI, 410–412
- Multidimensional rating scales  
BPRS, 365–366  
PANSS, 368–369  
present state examination, 367–368
- Multidomain evaluation, 560
- Multiinformant assessment, 475–476
- Multimethod design, 204–205, 472–474
- Multimodal behavior, 465  
assessments, 436
- Multiple attributes, 465, 467t
- Multiple behavior problems, 465, 466t
- Multiple causality, 465, 467t
- Multiple dimensions, 466t
- Multiple Errands Task, 585
- Multiple instruments, 474–475
- Multiple relationships, 544
- Multiple response modes, 466t
- Multiple sampling strategies, 481–482
- Multiple sclerosis (MS), 579–580, 584
- Multitasking, computer-automated  
assessment of, 577
- Multivariate base rates, 133–134
- Mutism  
behaviors, 347  
selective, 347
- Mutuality of Autonomy (MOA), 429
- My Next Move, 181  
for Veterans, 181
- N**
- NAART. *See* North American Adult Reading Test (NAART)
- NAB. *See* Neuropsychological Assessment Battery (NAB)
- “Narrow band” checklists, 444
- NART. *See* National Adult Reading Test (NART)
- Nation, 554
- National Adult Reading Test (NART), 132, 563
- National Board of Medical Examiners (NBME), 151
- National Institute on Aging/Alzheimer’s Association, 513
- National Standards for Foreign Language Learning (NSFLL), 560–561
- Naturalistic behavioral observation method, 481–482
- NCE. *See* Normal curve equivalents (NCE)
- NCLB Act. *See* No Child Left Behind Act (NCLB Act)
- NDRT. *See* Nelson–Denny Reading Test (NDRT)
- Negative Treatment Indicators (TRT), 408–409, 409t
- Neglect. *See* Extinction
- Nelson–Denny Reading Test (NDRT), 159–160
- NEO Personality Inventory (NEO-PI), 398
- NEPSY-II, 446–447
- Neural networks, 582–583
- Neurobehavioral Evaluation System 2 (NES 2), 586
- Neurodevelopmental assessment model, 219
- Neuroimaging, 581–582
- Neurological disorders, 238–239
- Neuropsychological assessment (NP assessment), 12–13, 20, 71–72, 75, 227, 234, 289, 294, 522  
attention, 516–517  
executive functioning, 517  
learning and memory, 516  
motor skills, 517  
perception, 517  
praxis, 518  
principles of older adult, 515–518  
verbal skills, 517  
visuospatial organization, 518
- Neuropsychological Assessment Battery (NAB), 241
- Neuropsychological Deficit Scale, 244
- Neuropsychological evaluations, 193–194
- Neuropsychological test, 227–228, 240–241  
batteries  
issues relating to validity and reliability, 234–240  
practical concerns in test construction, 233–234  
reliability, 240–241  
special problems in construction and standardization, 232–241

- Neuropsychologists, 75, 200, 209–210, 235–236
- Neuropsychology, 276, 554  
abounds with constructs, 239  
assessments, 79  
laboratory, 242
- Neuroscience of Rorschach, 423–424
- Neurotic complex, 14
- No Child Left Behind Act (NCLB Act), 149
- Nominal scales, 48–49
- Nomothetic assessment strategies, 464
- Non-English-dominant bilinguals, 560
- Noncompliance, 442
- Nonconditional assessment strategies, 465–471
- Nondeclarative memory, 507
- Nonforced choice stand-alone measures, 215
- Nonoccupational scales, 181
- Nonverbal communication, 309, 559
- Nonword repetition, 92
- Norm-referenced score interpretations, 51–52
- Normal aging, 506–508
- Normal curve, 52
- Normal curve equivalents (NCE), 52
- Normative quality research, 425–426
- Normative scores, 46–47
- Normative Update (NU), 155
- Norming and profile report of CISS, 175
- Norms. *See* Normative scores
- North American Adult Reading Test (NAART), 132
- North-American psychological tradition, 554
- Northern-American standardized testing procedures, 555
- NP assessment. *See* Neuropsychological assessment (NP assessment)
- NSFLL. *See* National Standards for Foreign Language Learning (NSFLL)
- Nurses' Observation Scale for Inpatient Evaluation-30 (NOSIE-30), 8
- O**
- O\*NET, 173–174  
Career Exploration Tools, 181  
Interest Profiler, 181–183
- Object relations theory, 429
- Objective measures, 560
- Objective personality tests, 10–11
- Observable behavioral responses, 472–473
- “Observed depression” rating, 373–374
- Observers, 442, 452–453
- Obsessiveness scale (OBS), 409t
- Occasion variance, 317
- Occupational alternatives, 177
- Occupational Extroversion–Introversion Scale, 181
- Occupational Interest Inventory, 170
- Occupational Interest Profiles (OIP), 181–182
- Occupational Scales, 173–175, 178
- Occupations Finder, The*, 177
- ODD. *See* Oppositional defiant disorder (ODD)
- Off-field evaluations, 290–292
- Off-task, 442
- Office Practices Basic Scales, 174
- Oklahoma Premorbid Intelligence Estimate-3 (OPIE-3), 132
- “Omnibus” checklists, 443–444
- “On-field” assessment, 290
- Ontogenetic development, 194–195
- Operant conditioning, 438
- Open interview, 318–319
- Oppositional defiant disorder (ODD), 435–436
- Oral Dependency Language (ODL), 428–429
- Oral Language composite, 156
- Ordinal scales, 49
- Orectic processing, 66
- Organizing (Orientation Scale), 174
- Orientation Scales, 174
- Overt behaviors, 435, 440
- Overt–motor indications of anxiety, 472–473
- P**
- Pacers, 276
- PAI. *See* Personality Assessment Inventory (PAI)
- Pair Cancellation, 92
- PANSS. *See* Positive and negative syndrome scale (PANSS)
- Paper-and-pencil screening tests, 227–228
- Paper-and-pencil tests, 3–4

- Paralinguistic behaviors, 486  
Paranoia (*Pa*), 406, 407<sup>t</sup>  
“Paranoid-destructive” factor, 374–375  
Parenting Stress Index—Fourth Edition (PSI-4), 451  
PARiConnect, 177  
Pathognomic scales, 256  
Pathological responses of MOA (MAP), 429  
“Pattern of perceptive process”, 422  
PC. *See* Picture Completion (PC)  
PDE. *See* Personality Disorder Examination (PDE)  
PDS. *See* Personal Data Sheet (PDS)  
Peabody Individual Achievement Test, 246  
Pediatric intelligence tests  
    framework for interpreting intelligence tests, 70–71  
    individually administered tests of intelligence for youth, 82  
    qualitative-idiographic approaches, 71–72  
    qualitative-nomothetic approaches, 72–74  
    quantitative-idiographic approaches, 74–79  
    quantitative-nomothetic approaches, 79–82  
    WISC-V, 82–87  
    WJ-IV COG, 87–93  
Pediatric performance validity testing, 214  
Percent agreement, 57  
Percentage of minus form quality answers (FQ-%), 428  
Percentile rank, 52–53  
*Perceptanalysis* (Piotrowski), 420  
Perception, 229–230, 517  
Perceptual disorders, 246  
    subbattery, 249  
Perceptual Organization Index, 107–108  
Perceptual Reasoning factor, 107–108  
Perceptual Reasoning Index (PRI), 84, 113–115. *See also* Processing Speed Index (PSI)  
    BD, 114  
    FW, 114  
    MR, 114  
    PC, 114–115  
    VP, 113  
Perceptual-Reasoning domain, 110  
Perceptual–cognitive conceptualization of Rorschach, 422  
Performance IQ (PIQ), 107–108  
Performance validity testing, 215  
Personal Data Sheet (PDS), 398  
Personal Style Scales, 179  
Personality, 282–283  
    assessment, 6–11, 20–21  
        interview, 6–8  
        objective personality tests, 10–11  
        projective personality tests, 8–10  
    characteristics, 50  
    measure, 398  
    personality-based paradigms, 461–462  
    testing, 558  
Personality Assessment Inventory (PAI), 401, 411  
    PsycINFO publications on, 410–412, 410<sup>t</sup>  
Personality Disorder Examination (PDE), 361–362  
PET scan. *See* Positron Emission Tomography scan (PET scan)  
16PF. *See* Sixteen Personality Factors Questionnaire (16PF)  
Phobia disorder, 17, 342  
Phobic object, 14  
Phoneme–Grapheme Knowledge Cluster, 155–156  
Phonological processing, 92  
PHR/GPHR. *See* Poor human representations to undistorted human representation (PHR/GPHR)  
Physical aggression, 442  
PIAT-R/NU. *See* Individual achievement test-revised–normative update (PIAT-R/NU)  
Picture Arrangement and Object Assembly, 108, 110  
Picture Completion (PC), 113–115  
    subtest of the Wechsler scales, 232  
Picture recognition, 92  
PIQ. *See* Performance IQ (PIQ)  
Pk scale. *See* Post-traumatic stress disorder (Pk scale)  
Plan implementation, 60–61, 60<sup>t</sup>  
    implement plan, reevaluate, and modify test, 60–61  
    preparing test manual, 61  
    submitting test proposal, 60  
PMA. *See* Primary Math Assessment (PMA)  
PocketSCAT2, 291

- POMS. *See* Profile of Mood States (POMS)
- Poor human representations to undistorted human representation (PHR/GPHR), 428
- Poor test session behavior, 74
- POR. *See* Problem-oriented record (POR)
- Positive and negative syndrome scale (PANSS), 368–369
- Positive predictive power (PPP), 219
- Positron Emission Tomography scan (PET scan), 234–235
- Post-traumatic stress disorder (Pk scale), 408t
- Posthumous paper, 420, 422
- Potential value of theory, 205–206
- Poverty, 211–213
- PPI. *See* Psychological Performance Inventory (PPI)
- PPP. *See* Positive predictive power (PPP)
- Practitioners, 279
- Praxis, 518
- Precursors to self-report inventories of psychopathology, 397–398
- Prediction, Prescription and Process (Three Ps), 70–71
- Predictive validity, 237–238
- Present State Examination (PSE), 360, 367–368
- PRI. *See* Perceptual Reasoning Index (PRI)
- Primary Math Assessment (PMA), 162
- Primary memory, 228–229
- Primary-mental abilities, 206
- Problem-oriented record (POR), 13, 15
- Problematrical interpersonal function, 428
- Processing Speed, 107–110
- Processing Speed Index (PSI), 85, 116–118. *See also* Perceptual Reasoning Index (PRI)
- cancellation, 117–118
  - coding, 117
  - symbol search, 117
- Producing (Orientation Scale), 174
- Professional schools, 144–146
- Profile
- analysis, 518
  - matching, 242
- Profile Elevation and Impairment scales, 256–257
- Profile of Mood States (POMS), 283
- Projective methods, 422
- Projective personality tests, 8–10
- Prosody–Face Matching, Social Cognition, 129
- Prosody–Pair Matching, Social Cognition, 129
- Protocol validity assessment with MMPI-2, 404–406
- PSE. *See* Present State Examination (PSE)
- Pseudodementia, 512
- PSI. *See* Processing Speed Index (PSI)
- PSI-4. *See* Parenting Stress Index—Fourth Edition (PSI-4)
- Psychasthenia (*Pt*), 32–33, 406, 407t
- Psychiatric/psychiatry, 307
- conditions, 510–512
  - depression, 510–512
  - diagnosis, 11, 355–356
  - measure of psychiatric disability, 250
- Psychoanalytic concepts, 437–438
- Psychodiagnostik*, 420–422
- Psychoeducational
- assessments, 200
  - evaluation, 199–201
- Psychological assessment, 70–71, 463–464
- behavioral assessment, 13–19
  - clinical assessments, 508–512
  - depression, 510–512
  - family history, 509
  - medical conditions, 508–509
  - psychiatric conditions, 510–512
  - schizophrenia, 512
  - social adaptation, 509
- cognitive functioning, 512–515
- developments, 19–21
- differential diagnosis
- AD, 519
  - depression *vs.* dementia, 520–521, 521t
  - profile analysis, 518
  - vascular dementia, 520
- of elderly, 505
- intelligence and achievement testing, 4–5
- measurement and clinical science, 461–463
- idiographic and nomothetic assessment strategies, 464
- integrating multiple measures in clinical assessment, 488–489
- multiinformant assessment, 475–476

- multimethod assessment, 472–474  
multiple instruments, 474–475  
time-series and repeated measurement, 476–477  
neuropsychological assessment, 12–13  
    principles of older adult, 515–518  
normal aging, 506–508  
personality assessment, 6–11  
special problems in geriatric assessments, 522  
    in sport, 275  
Psychological functions, higher, 196–198  
Psychological Performance Inventory (PPI), 286  
Psychological tests, 48, 425–426, 539–540, 555  
    bias in, 551–552  
    conceptualization, 32–34, 32*t*  
    fairness in, 551–553  
    listing, 538  
    plan implementation, 60–61, 60*t*  
        implement plan, reevaluate, and modify test, 60–61  
        preparing test manual, 61  
        submitting test proposal, 60  
    planning standardization, scaling, and psychometric studies, 46–60, 46*t*  
    specifications  
        age range, 35  
        developing TOSs or test blueprint, 37–39  
        internal structure, 35*t*, 36–37  
        item estimation, 40–41  
        item format, 39–40  
        methods for item tryout and selection, 42–46  
        plan for item development, 41–42  
        test format, 35–36, 35*t*  
Psychological/psychology, 65–66, 307  
    construct, 555–557  
    evaluation, 19  
    factor, 369–370  
    history, 533  
Psychologists, 196, 551–552  
    language proficiency and familiarity, 560–561  
Psychometric(s), 3–4  
    factors, 133  
    information, 341  
    matched test, 556  
    studies, 46–60, 46*t*  
Psychopathic Deviation (*Pd*), 406, 407*t*  
Psychopathology assessment, 11  
    MMPI-2  
        clinical scales, 406, 407*t*  
        content scales, 408–409  
        supplementary scales, 407–408  
Psychophysiological assessment method, 486–487  
Psychosis, 8  
Psychostimulant medications, 216  
Psychotherapy, 82  
Psychotic episode, 363  
Psychotic symptoms, 375–380  
    BNSS, 379–380  
    SANS, 378  
    SAPS, 377–378  
PsycINFO database, 410  
*Pt.* *See* Psychasthenia (*Pt*)  
Public domain item pool for personality assessment, 182  
Public schools, achievement testing in, 149–150  
Purdue Interest Report, 170
- Q**
- Q-interactive assessment, 121  
Q-interactive platform, 135  
QEd. *See* Educational quality measure (QEd)  
QoL-AD. *See* Alzheimer's Disease-related Quality of Life scale (QoL-AD)  
Qualitative interpretation, 70  
Qualitative linguistic information, 561  
Qualitative-idiographic approaches, 71–72  
Qualitative-nomothetic approaches, 72–74  
Quality of education, 4, 20, 155, 211  
Quantitative indices, 259  
Quantitative Study of Frontal Lobes, 242  
Quantitative-idiographic approaches, 74–79  
Quantitative-nomothetic approaches, 79–82  
Quick Inventory of Depressive Symptomatology (QIDS), 372
- R**
- R-optimized method, 427  
R-PAS. *See* Rorschach Performance Assessment System (R-PAS)  
R-type factor analysis, 240

- Race, 211, 554  
 Racial/ethnic norms, 557  
 Rapid Marital Interaction Coding System, 482–483  
 Rapport, 279  
**RAS.** *See* Restricted Academic Situation (RAS)  
 Rating scales, 169–170  
     multidimensional rating scales, 365–369  
     symptom and behavior, 362–380  
     symptom specific and clinical diagnosis  
         affective symptoms, 369–375  
         psychotic symptoms, 375–380  
 Ratio scales, 50–55  
 Rational sampling approach, 170  
 Rational-based item selection method, 399<sub>t</sub>, 404–406  
 Raven’s Progressive Matrices, 105  
 Raw scores, 51  
**RC.** *See* Reliable change (RC)  
**RC Scales.** *See* Restructured Clinical Scales (RC Scales)  
**RCI.** *See* Reliable change index (RCI)  
**RCMAS-2.** *See* Revised Children’s Manifest Anxiety Scale: Second Edition (RCMAS-2)  
**RDC.** *See* Research Diagnostic Criteria (RDC)  
**RDoC.** *See* Research Domain Criteria (RDoC)  
**RDS.** *See* Reliable digit span (RDS)  
 Re-learning motoric behavior, 277  
 Reading  
     cognitive processes involving in reading achievement, 158–159  
     comprehension measures, 159–161  
     disorders, 159  
     level, 563  
     scales, 256  
 Receptive Speech scales, 256, 258  
 Recommendations, 203  
     in forensic psychology assessment, 538–539  
 Record keeping, 533  
     in forensic psychology, 545–546  
**REDSOCS.** *See* Revised Edition of School Observation Coding System (REDSOCS)  
 Referral information, 537  
 Referral questions, 199–200  
**Reitan Aphasia Screening test**, 243–247, 249  
 Reitan’s program, 247–248  
**Relevance 2050 Initiative**, 554, 563  
 “Relevance for treatment”, 437  
 Reliability, 55, 175–177, 179, 181, 240–241, 317–318  
     coefficients, 55  
     issues relating to, 234–240  
     item formats promoting, 40  
     of psychological test results, 538  
     reliability/precision studies, 55–58  
 Reliable change (RC), 293–294  
 Reliable change index (RCI), 76–77  
 Reliable digit span (RDS), 215–216  
 Reliable scores, 60  
 Remote behavioral assessment, 579–580  
 Remote psychological assessment, 579–581  
 Renard Diagnostic Interview, 6–7  
 Reporting, 174  
 Repression, 473  
 Research design and methodology, 203–204  
     criteria for judging effectiveness, 204  
 Research Diagnostic Criteria (RDC), 310–314, 321  
 Research Domain Criteria (RDoC), 582  
 Researchers, 279, 282  
 Resilience, 285  
 Response(s)  
     bias, 215  
     contingencies, 467  
     desynchrony, 473  
     to tests, 552  
**Restricted Academic Situation (RAS)**, 442–443  
**Restructured Clinical Scales (RC Scales)**, 399<sub>t</sub>, 401–402  
**Revised Children’s Manifest Anxiety Scale: Second Edition (RCMAS-2)**, 447  
**Revised Conflict Tactics Scale (CTS2)**, 485  
**Revised Edition of School Observation Coding System (REDSOCS)**, 442  
 Rey Auditory-Verbal Learning Test, 519  
 Rey Complex Figure Test, 242  
 Reynolds Intellectual Assessment Scales-2, 105  
 Rey-Osterrieth Complex Figure Test, 229–230

- Rhythm scales, 256  
Rhythmic patterns, 230–231  
Right Hemisphere scales, 256  
Risk-Taking scale, 179  
Rohling interpretive method, 242  
Rorschach human movement responses, 424  
Rorschach Oral Dependency (ROD).  
    *See* Rorschach Oral Dependency  
    Language scale  
Rorschach Oral Dependency Language scale, 428–429  
Rorschach Performance Assessment System (R-PAS), 423, 425–427, 429  
Rorschach research  
    history and development, 419–423  
    neuroscience, 423–424  
    new clinical developments, 426–429  
        complexity, 427–428  
        Ego Impairment Index, 428  
        MOA, 429  
        ODL, 428–429  
        R-optimized method, 427  
    normative and form quality research, 425–426  
    Rorschach variable selection, 424–425  
Rorschach technique, 8–9, 12–13, 420, 461–462
- S**
- SAC. *See* Standardized Assessment of Concussion (SAC)  
SADS. *See* Schedule for Affective Disorders and Schizophrenia (SADS)  
SAINT-II system, 574  
Sales, 174  
SASC-R. *See* Social Anxiety Scale for Children-Revised (SASC-R)  
SAT. *See* Scholastic Aptitude Test (SAT)  
SBS Inventory of Social Behavior Standards and Expectations, 452  
SC. *See* Self-Control (SC)  
Scale for Assessment of Negative Symptoms (SANS), 377–378  
Scale for Assessment of Positive Symptoms (SAPS), 377–378  
Scale(s), 48  
    for CISS, 174  
    construction methods, 185, 399t, 403, 408–409  
development procedures, 402  
of measurement, 48–55  
Scaled scores, 51–52  
Scaling method, 46–60, 46t  
SCAT. *See* Sport Concussion Assessment Tool (SCAT)  
Scenario-based assessment, 576–578  
Schedule for Affective Disorders and Schizophrenia (SADS), 6–7, 370  
    for school-aged children, 340–341  
*Schedule for Affective Disorders And Schizophrenia For School-Aged Children Present And Lifetime Version* (K-SADS-PL), 340–341  
Schedules for Clinical Assessment in Neuropsychiatry (SCAN), 367–368  
Scheduling addresses, 315  
Schizophrenia (*Sc*), 253–254, 357–358, 406, 407t, 512  
Scholastic Aptitude Test (SAT), 144  
School Situations Questionnaire (SSQ), 450  
School Situations Questionnaire and revision (SSQ-R), 450  
School-aged children, SADS for, 340–341  
SCICA. *See* Semistructured Clinical Interview for Children and Adolescents (SCICA)  
SCID. *See* Structured Clinical Interview for DSM-III (SCID)  
SCID-5 Alternative Model for Personality Disorders (SCID-5-AMPD), 361  
SCID-5 Clinical Version (SCID-5-CV), 357  
SCID-5 Research Version (SCID-5-RV), 357  
SCID-5-PD. *See* Structured Clinical Interview for the DSM-5 Personality Disorders (SCID-5-PD)  
SCID-5-SPQ. *See* Structured Clinical Interview for DSM-5 Screening Personality Questionnaire (SCID-5-SPQ)  
SCID-5. *See* Structured Clinical Interview for DSM-5 (SCID-5)  
Science, 174  
Scoring format, 174  
SDS. *See* Self-Directed Search (SDS)  
Seashore Rhythm test, 243, 245  
Seashore Tonal Memory Test, 246  
Second-order abilities, 206, 207t  
SED. *See* Standard error of difference (SED)

- Selection process, 103  
*Selective Mutism Comprehensive Diagnostic Questionnaire*, 347  
Self-Control (SC), 176  
Self-Directed Search (SDS), 173, 176  
Self-monitoring, 448–449, 583–585  
    method, 483–484  
Self-regulatory systems and plans, 439  
Self-report, 317  
    instruments, 447–448  
    inventories, 398  
        of psychopathology, 397–398  
    method, 472, 474  
Selz–Reitan rules, 243–244  
SEM. *See* Standard errors of measurement (SEM)  
Semistructured Clinical Interview for Children and Adolescents (SCICA), 445  
Semistructured interviews, 321–323, 339–340, 348  
    for children and adolescents, 340–341  
SENAS. *See* Spanish and English Neuropsychological Assessment Scales (SENAS)  
Sensory–perceptual abilities, 206, 207  
Sentence Comprehension subtest, 154  
Sequential system of construct-oriented scale development, 399<sub>t</sub>, 402  
Sequin–Goddard Formboard, 245  
Serial assessment with WAIS-IV and WMS-IV, 133  
SES. *See* Socioeconomic status (SES)  
*Si.* *See* Social Introversion (*Si*)  
Sideline evaluations, 290–295  
SII. *See* Strong Interest Inventory (SII)  
Similarities subtest (SI subtest), 111–112  
Simulated observations, 442–443  
Simulators, 586  
Single Photon Emission Computerized Tomography (SPECT), 234–235  
Sixteen Personality Factors Questionnaire (16PF), 398  
Ski Instructor criterion samples, 175  
Smartphones for psychological assessment, 584–585  
Social adaptation, 509  
Social Anxiety Scale for Children-Revised (SASC-R), 447  
Social desirability  
    bias, 279  
    scales, 37  
Social discomfort scale (SOD), 409<sub>t</sub>  
Social ecology, 436–437  
Social interactions, 196  
Social Introversion (*Si*), 406, 407<sub>t</sub>  
Social learning  
    theorists, 438–439  
    theory, 435, 440  
Social network, membership in, 507  
Social Security Administration, 545  
Social Skill Behavioral Assessment System, 482–483  
Social Skills Rating System (SSRS), 444  
Social stressors, 464  
Social Support Appraisals Scale (APP), 452  
Social Support Scale for Children (SPPC), 452  
Socialization model, 284–285  
Socially sensitive behaviors, 482  
Society for Sport, Exercise, and Performance Psychology, 276  
Socioeconomic status (SES), 211, 275, 557, 562–563  
    traditional measures, 562  
Sociopolitical educational environment, 143  
Solicitation of teacher, 442  
Somatic  
    factor, 369–370  
    hallucination, 377  
Somatization, 366  
Sources of error, 209–219  
    clinical decision making, 216–219  
    culture, language, and poverty, 211–213  
    ecological validity, 213  
    incremental validity, 209–211  
    malingering, 214–216  
Spaceflight Cognitive Assessment Tool for Windows (WinSCAT), 580–581  
Spanish and English Neuropsychological Assessment Scales (SENAS), 556  
Spatial Addition, 124  
Specialty Guidelines for Forensic Psychology, 539, 543–544  
Specific disorders type of measure, 329  
SPECT. *See* Single Photon Emission Computerized Tomography (SPECT)

- Speech  
  neuropsychological assessment, 231  
  perception test, 243, 245
- Speed impact on athletic prowess, 275
- Split-half method, 240–241
- Sport Concussion Assessment Tool (SCAT), 289
- Sport-related concussions (SRC), 276, 288
- Sports  
  constructs and behaviors measuring in assessments, 280–281, 281t  
  milieu, 277–278  
  neuropsychology, 276–277, 287  
  programs quality, 275  
  psychology disciplines, 276–277  
  psychology, 277–280, 280t
- Sports Neuropsychology Society, 287–288
- SPPC. *See* Social Support Scale for Children (SPPC)
- SSQ. *See* School Situations Questionnaire (SSQ)
- SSQ-R. *See* School Situations Questionnaire and revision (SSQ-R)
- SSRS. *See* Social Skills Rating System (SSRS)
- Stability of interests, 182
- STAIC. *See* State-Trait Anxiety Inventory for Children (STAIC)
- Stamina, physical attribute, 275
- Standard error of difference (SED), 210
- Standard errors of measurement (SEM), 210
- Standard scores, 51–52, 81
- Standard Version of Luria's Neuropsychological Tests, 255
- Standardization  
  addresses, 315  
  planning, 46–60, 46t  
    specifying reliability/precision studies, 55–58  
    specifying scaling methods, 48–55  
    specifying standardization plan, 46–48  
    specifying validity studies, 58–60  
  in test administration, 575
- Standardized Assessment of Concussion (SAC), 289
- Standards for Educational and Psychological Testing, 58
- Standards-based interpretations, 53–54
- Stanford Center on Longevity, 583–584
- Stanford–Binet Intelligence Scale.  
  *See* Binet–Simon Scale
- State-Trait Anxiety Inventory for Children (STAIC), 447
- Statistical analyses, 170, 481
- Statistical (Orientation Scale), 174
- Stewardess*, 171
- Stimulus-to-fantasy approach, 422
- Story Memory Test, 246
- Strength, physical attribute, 275
- Stress, 473  
  stress–headache relationship, 473–474
- Strong Interest Inventory (SII), 170–173, 177–179, 185
- Strong Vocational Interest Blank, 171, 177–178
- Stroop, 213
- Structured Clinical Interview for DSM-5 (SCID-5), 356–358
- Structured Clinical Interview for DSM-5 Screening Personality Questionnaire (SCID-5-SPQ), 361
- Structured Clinical Interview for DSM-III (SCID), 7, 355
- Structured Clinical Interview for DSM-IV Childhood Diagnoses* (KID-SCID), 344–345
- Structured Clinical Interview for the DSM-5 Personality Disorders (SCID-5-PD), 356, 361–362
- Structured interviews, 321–323, 339–348.  
  *See also* Clinical interview;  
  Diagnostic interviews  
  anxiety disorders, 345–346  
  areas, 347–348  
  CAPA, 342–343  
  ChIPS, 343  
  diagnostic interview for children, 341–342  
  DICA, 341  
  MINI for children and adolescents, 343–344  
  SADS for school-aged children, 340–341  
  selective mutism, 347  
  strengths and limitations, 348–349  
  structured clinical interview for DSM-IV  
    childhood diagnoses, 344–345  
    trauma-related conditions, 346–347
- Subject variance, 317

- Subject-sampling strategy, 481–482  
Subjective values, 439  
Suboptimal effort, 215  
Subpoenas  
  dealing with, 546  
  responding to, 533  
Substance use disorder diagnosis, 317  
Subtest level changes, 108–111  
  cancellation, 111  
Figure Weights, 110  
perceptual reasoning, 109  
Processing Speed, 109–110  
Verbal Comprehension, 108–109  
Visual Puzzles, 110  
Working Memory, 109  
Subtest-score differences, 210  
Supervision, 174  
Supplementary Scales, 407–408  
Symbol search, 117  
Symbol Span, 124  
Symptom Validity Scale, 404–406  
Symptom validity testing, 214  
Symptom-reported validity tests, 403  
Synthetic work (SynWork), 577  
Systematic observation of causal  
  variable–target behavior functional  
  relations, 491  
Systematizers, 421  
Systemic variance, 204
- T**  
T-score profile, 244–247  
Table of Specifications (TOSSs), 37–39  
  development, 37–39  
  sample, 38<sup>t</sup>  
Tactile Hallucination, 377  
Tactile neglect, 230–231  
Tactile scales, 256, 260–261  
Tactual Performance test, 243, 249  
Target behaviors, 472–473  
  operationally definitions of, 489–490  
Task-based assessment, 576–578  
TAT. *See* Thematic Apperception Test  
  (TAT)  
TBI. *See* Traumatic brain injury (TBI)  
Teacher-treatment compatibility, 450–451  
Team  
  cohesiveness, 286–287  
  managers, 282  
orientation scale, 179  
Technological developments in assessment,  
  573–576  
access to care and telehealth, 579–581  
cognitive rehabilitating and self-  
  monitoring, 583–585  
computers for cognitive training, 584  
enhancing diagnosis and behavioral  
  prediction, 582–583  
enhancing efficiency and reliability,  
  574–576  
advantages and challenges in early  
  adoption, 575  
  technological advances, 575–576  
expanding research options, 586–587  
expanding tasks and scenario-based  
  assessment, 576–578  
linking cognitive domains and biological  
  systems, 581–582  
Telehealth, 579–581  
Telemedicine, 579  
Telemental health, 579  
Telepsychiatry, 579  
Test blueprint development, 37–39  
Test conceptualization, 32–34, 32<sup>t</sup>  
  applications, 33  
  specify conceptual and operational  
    definitions of constructs, 34  
  specify users, 34  
Test content, 552  
  evidence based on, 58  
Test format and structure specification,  
  35–46, 35<sup>t</sup>  
  developing TOSSs or test blueprint, 37–39  
  item estimation, 40–41  
  plan for item development, 41–42  
  specify age range, 35  
  specify internal structure, 36–37  
  specify item format, 39–40  
  specify methods for item tryout and  
    selection, 42–46  
    diversity panel, 42  
    specifying plan for item tryout, 42  
    specifying statistical methods, 42–46  
  specify test format, 35–36  
Test information function (TIF), 57–58  
Test interpretation, 198  
Test manuals, 185  
  preparation, 61

- Test norms, 557–558  
Test Observation Form, 72–73  
Test of Premorbid Functioning (TOPF),  
    130–132  
Test Pilot criterion samples, 175  
Test proposal submission, 60  
Test translation and adaptation,  
    555–556  
Test-battery design, 198  
Test-taker, 552  
    level of acculturation, 558–559  
Test-taking skills, 211  
Testimony  
    courtroom, 541–542, 541t  
    direct, 541–542, 542t  
“Testing for brain damage”, 227  
“Testing for organicity”, 227  
Test-retest  
    coefficients, 56  
    method, 240–241  
    reliability, 344, 358  
Teuber group, 236  
Thematic Apperception Test (TAT), 8–9,  
    461–462  
Theory of careers (Holland), 176  
Third-order clusters of abilities, 206, 207t  
Third-party  
    individuals and/or records, 539  
    involvement, obligations, or entitlements,  
        538  
    observer effects, 538  
Three Ps. *See* Prediction, Prescription and  
    Process (Three Ps)  
Three-stratum theory (Carroll), 205–206  
TIF. *See* Test information function (TIF)  
Time Sense test, 243, 248–249  
Time-sampling  
    error, 55  
    strategy, 481–482, 484  
Time-series and repeated measurement,  
    476–477  
Timeline follow-back interview procedure,  
    486  
TOPF. *See* Test of Premorbid Functioning  
    (TOPF)  
TOSs. *See* Table of Specifications (TOSs)  
Tower of London, 213  
Tower of London—Drexel University  
    (ToLDx), 556  
Traditional “paper and pencil”  
    neuropsychological tests, 292  
Trail Making test, 242–243, 246  
Transdiagnostic constructs, 342  
Transient variables function, 427–428  
Trauma-related conditions, 346–347  
Traumatic Aphasia (1970), 255  
Traumatic brain injury (TBI), 584  
Triplett, 276  
TRT. *See* Negative Treatment Indicators  
    (TRT)  
Try-out activities, 169–170  
Type A scale (TPA scale), 408–409, 409t  
Type–locus interaction, 242–243
- U**
- U.S. legal decisions on scientific expertise,  
    536–537  
*UCLA Child/Adolescent PTSD Reaction  
    Index for DSM-5*, 346–347  
“Unconditional positive regard”, 310–314  
Undergraduate grade point averages  
    (UGPA), 145–146  
United States Medical Licensure  
    Examination (USMLE), 145–146  
University of Michigan version of CIDI  
    (UM-CIDI), 360  
University of Minnesota, 401–404  
Unsolved problems, 449–450
- V**
- VA. *See* Veterans affairs (VA)  
Validity, 175–177, 179, 181, 317–318, 470t  
    issues relating to, 234–240  
    item formats promoting, 40  
    of psychological test results, 538  
    studies, 58–60  
        evidence based on relations to  
            variables, 58–60  
        evidence based on test content, 58  
            tests, 214–215  
Validity scales. *See* Dissimulation scales  
Variables, evidence based on relations to,  
    58–60  
Variance, 204  
Vascular dementia, 520  
VC subtest. *See* Vocabulary subtest  
    (VC subtest)  
Verbal communication, 309, 559

- Verbal communication (*Continued*)  
     factor, 107–108
- Verbal Comprehension, 108–109
- Verbal Comprehension Index (VCI), 85, 107–108, 111–113  
     comprehension, 112–113  
     information, 112  
     similarities, 112  
     vocabulary, 112
- Verbal IQ (VIQ), 107–108
- Verbal Paired Associates, 125
- Verbal skills, 517
- Veridicality, 213
- Verisimilitude, 213
- VEs. *See* Virtual environments (VEs)
- Veterans affairs (VA), 10
- Veterans and Military Occupations Finder, The*, 177
- VII. *See* Vocational Interest Inventory (VII)
- VIQ. *See* Verbal IQ (VIQ)
- Virtual environments (VEs), 578  
     for ecologically valid assessments, 577–578
- Virtual patients (VPs), 326
- Virtual reality equipment, 578
- Visual field examination, 246
- Visual imagination, 419–420
- Visual Memory, 125–126  
     Designs, 126  
     Visual Reproduction, 126
- Visual Memory Index (VMI), 125
- Visual Puzzles (VP), 85, 110, 113
- Visual Reproduction, 126
- Visual scales, 256
- Visual Spatial Index (VSI), 85, 107–108
- Visual stimuli, 419
- Visual Working Memory, 123–124  
     Spatial Addition, 124  
     Symbol Span, 124
- Visual Working Memory Index (VWMI), 123
- Visualization, 92
- Visual–spatial processing, 113
- Visual–spatial skills, 230
- Visuospatial, 276–277  
     organization, 518
- VMI. *See* Visual Memory Index (VMI)
- Vocabulary subtest (VC subtest), 111–112
- Vocational interest
- blank, 170  
     of Men and Women, 169  
     theories, 172–173  
         construction of interest inventory scales, 172–173
- Vocational Interest Inventory (VII), 173
- Vocational Preference Inventory (VPI), 176
- Vocational preparation and employment, 146–147
- Voir Dire, 536, 541, 541*t*
- VP. *See* Visual Puzzles (VP)
- VPs. *See* Virtual patients (VPs)
- VSI. *See* Visual Spatial Index (VSI)
- Vulnerable abilities, 206, 207*t*
- VWMI. *See* Visual Working Memory Index (VWMI)
- Vygotsky’s theories, 195–196
- W**
- WCST. *See* Wisconsin Card Sorting Test (WCST)
- Web-based administration, 174
- Wechsler Adult Intelligence Scale (WAIS), 242
- Wechsler Adult Intelligence Scale—fourth edition (WAIS-IV), 105–106, 516  
     development approach, 106–108  
     and digital assessment, 121–122  
     index scores and structure, 111–119  
         CPI, 119  
         GAI, 118–119  
         PRI, 113–115  
         PSI, 116–118  
         VCI, 111–113  
         WMI, 115–116  
     joint factor structure of WMS-IV and, 128–129  
     refining interpretation, 130–135  
         cognitive variability, 134–135  
         demographic referenced norms, 131  
         multivariate base rates, 133–134  
         serial assessment with WAIS-IV and WMS-IV, 133  
         TOPF, 131–132  
     serial assessment with WMS-IV and, 133
- Wechsler Adult Intelligence Scale—fourth edition/Wechsler Memory Scale—fourth edition conormed battery (WAIS-IV/WMS-IV), 133

- Wechsler Adult intelligence Scale—third edition (WAIS-III), 106
- Wechsler Adult Intelligence Scale—third edition/Wechsler Memory Scale—third edition conormed battery (WAIS-III/WMS-III), 133
- Wechsler Individual Achievement Test-III (WIAT-III), 156–157
- Wechsler Intelligence Scale for Children—fifth edition (WISC-V), 82, 84, 86–87, 120, 210, 446–447  
additional interpretation, 86  
critique, 86–87  
properties, 84–85  
reliable change on, 87<sup>t</sup>  
standardization, 83–84  
subscales and indices, 78<sup>t</sup>  
theory, 82–83
- Wechsler Intelligence Scale for Children—fourth edition (WISC-IV), 107, 111
- Wechsler Intelligence Scales, 243, 245–246
- Wechsler Intelligence Scales for Children (WISC), 68–69, 72–73
- Wechsler Memory Scale (WMS), 12–13, 507
- Wechsler Memory Scale-III (WMS-III), 239
- Wechsler Memory Scale—fourth edition (WMS-IV), 105, 111<sup>f</sup>, 122–128  
Auditory Memory, 124–125  
BCSE, 127–128  
Delayed Memory, 127  
IMI, 126–127  
joint factor structure of WAIS-IV and, 128–129  
serial assessment with WAIS-IV and, 133  
Visual Memory, 125–126  
Visual Working Memory, 123–124
- Wechsler scales, 12–13  
scores, 52
- Wechsler Test of Adult Reading (WTAR), 132, 514
- Wechsler–Bellevue Intelligence Scale, 105
- Wechsler–Bellevue laterality studies, 254
- Wechsler–Bellevue studies of brain lesion lateralization, 236
- Weighted sum of cognitive codes (WSumCog), 428
- Western Aphasia Battery, 231–232
- WHO. *See* World Health Organization (WHO)
- WHODAS 2.0. *See* World Health Organization Disability Assessment Schedule, Version 2.0 (WHODAS 2.0)
- WIAT-III. *See* Wechsler Individual Achievement Test-III (WIAT-III)
- Wide Range Achievement Test-Revised (WRAT-R), 151, 246
- Wide Range Achievement Test—fourth edition (WRAT-4), 154–155, 563
- Wide-aperture attention, 232
- Wiesen Test of Mechanical Aptitude (WTMA), 147
- WinSCAT. *See* Spaceflight Cognitive Assessment Tool for Windows (WinSCAT)
- WISC. *See* Wechsler Intelligence Scales for Children (WISC)
- Wisconsin Card Sorting Test (WCST), 213, 246, 576–578
- WMI. *See* Working Memory Index (WMI)
- WMS. *See* Wechsler Memory Scale (WMS)
- WMS-III. *See* Wechsler Memory Scale-III (WMS-III)
- WMS-IV. *See* Wechsler Memory Scale—fourth edition (WMS-IV)
- Woodcock Munoz Language Survey-Revised, 560
- Woodcock–Johnson Psychoeducational Battery (WJ Psychoeducational Battery), 151
- Woodcock–Johnson Tests of Cognitive Abilities—fourth edition (WJ-IV COG), 82, 88–89, 92–93, 105, 155–156  
critique, 92–93  
properties, 89–93  
standardization, 88  
subscales and indices, 90<sup>t</sup>  
theory, 87–88
- Woodworking, 174
- Word Choice subtest, 130
- Word-reading test in English, 563
- Work Importance Locator, The, 181
- Work Interference (WRK), 408–409, 409<sup>t</sup>
- Work Style scale, 179
- Working Memory, 109

- Working Memory (*Continued*)  
factor, 107–108
- Working Memory Index (WMI), 85, 107–108, 115–116  
Arithmetic, 116  
Digit Span, 115  
Letter–Number Sequencing, 116
- World Health Organization (WHO), 586  
WHO/UCLA Test Battery, 556
- World Health Organization Disability Assessment Schedule, Version 2.0 (WHODAS 2.0), 329
- WRAT-4. *See* Wide Range Achievement Test—fourth edition (WRAT-4)
- WRAT-R. *See* Wide Range Achievement Test-Revised (WRAT-R)
- Writing scales, 256
- WSumCog. *See* Weighted sum of cognitive codes (WSumCog)
- WTAR. *See* Wechsler Test of Adult Reading (WTAR)
- WTMA. *See* Wiesen Test of Mechanical Aptitude (WTMA)
- Y**
- Young Mania Rating Scale (YMRS), 327, 371, 374–375
- Youth Self-Report (YSR), 447

# HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

Fourth Edition

Edited by **Gerald Goldstein<sup>(†)</sup>, Daniel N. Allen, and John DeLuca**

The fourth edition of the *Handbook of Psychological Assessment* provides scholarly overviews of the major areas of psychological assessment, including test development, psychometrics, technology of testing, and commonly used assessment measures. Psychological assessment is included for all ages with new coverage encompassing ethnic minorities and the elderly. Assessment methodology discussed includes formal testing, interviewing, and observation of behavior. The handbook also discusses assessment of personality and behavior including intelligence, aptitude, interest, achievement, and psychopathology. New coverage includes use of assessments in forensic applications.

## Key Features

- Encompasses test development, psychometrics, and assessment measures
- Covers assessment for all age groups
- Includes formal testing, interviews, and behavioral observation as testing measures
- Details assessments for intelligence, aptitude, achievement, personality, and psychopathology
- New coverage of assessment used in forensic psychology
- New coverage on assessments with ethnic minorities



ACADEMIC PRESS

An imprint of Elsevier  
[elsevier.com/books-and-journals](http://elsevier.com/books-and-journals)

ISBN 978-0-12-802203-0

A standard barcode for the ISBN 978-0-12-802203-0, with the numbers 9 780128 022030 printed below it.