THE GUILFORD PRACTICAL INTERVENTION IN THE SCHOOLS SERIES

Assessing Intelligence in Children and Adolescents

A Practical Guide

John H. Kranzler
Randy G. Floyd

ebook

THE GUILFORD PRESS

Assessing Intelligence in Children and Adolescents

The Guilford Practical Intervention in the Schools Series

Kenneth W. Merrell, Founding Editor
T. Chris Riley-Tillman, Series Editor

www.guilford.com/practical

This series presents the most reader-friendly resources available in key areas of evidence-based practice in school settings. Practitioners will find trustworthy guides on effective behavioral, mental health, and academic interventions, and assessment and measurement approaches. Covering all aspects of planning, implementing, and evaluating high-quality services for students, books in the series are carefully crafted for everyday utility. Features include ready-to-use reproducibles, lay-flat binding to facilitate photocopying, appealing visual elements, and an oversized format. Recent titles have Web pages where purchasers can download and print the reproducible materials.

RECENT VOLUMES

Cognitive Therapy for Adolescents in School Settings
Torrey A. Creed, Jarrod Reisweber, and Aaron T. Beck

Motivational Interviewing for Effective Classroom Management:
The Classroom Check-Up
Wendy M. Reinke, Keith C. Herman, and Randy Sprick

Positive Behavior Support in Secondary Schools: A Practical Guide
Ellie L. Young, Paul Caldarella, Michael J. Richardson, and K. Richard Young

Academic and Behavior Supports for At-Risk Students: Tier 2 Interventions
Melissa Stormont, Wendy M. Reinke, Keith C. Herman, and Erica S. Lembke

RTI Applications, Volume 1: Academic and Behavioral Interventions
Matthew K. Burns, T. Chris Riley-Tillman, and Amanda M. VanDerHeyden

Coaching Students with Executive Skills Deficits
Peg Dawson and Richard Guare

Enhancing Instructional Problem Solving:
An Efficient System for Assisting Struggling Learners
John C. Begeny, Ann C. Schulte, and Kent Johnson

Clinical Interviews for Children and Adolescents, Second Edition:
Assessment to Intervention
Stephanie H. McConaughy

RTI Team Building: Effective Collaboration and Data-Based Decision Making
Kelly Broxterman and Angela J. Whalen

RTI Applications, Volume 2: Assessment, Analysis, and Decision Making
T. Chris Riley-Tillman, Matthew K. Burns, and Kimberly Gibbons

Daily Behavior Report Cards: An Evidence-Based System of Assessment and Intervention
Robert J. Volpe and Gregory A. Fabiano

Assessing Intelligence in Children and Adolescents: A Practical Guide
John H. Kranzler and Randy G. Floyd

The RTI Approach to Evaluating Learning Disabilities
Joseph F. Kovaleski, Amanda M. VanDerHayden, and Edward S. Shapiro

Assessing Intelligence in Children and Adolescents

A Practical Guide

JOHN H. KRANZLER
RANDY G. FLOYD

THE GUILFORD PRESS
New York London

A Division of Guilford Publications, Inc.
72 Spring Street, New York, NY 10012
www.guilford.com

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data is available from the publisher.

ISBN 978-1-4625-1121-1

About the Authors

John H. Kranzler, PhD, is Professor and Program Director of School Psychology in the School of Special Education, School Psychology, and Early Childhood Studies at the University of Florida. He has received numerous awards for his research, including article-of-the-year awards in *School Psychology Quarterly* and *School Psychology Review*. Dr. Kranzler was Associate Editor of *School Psychology Quarterly* for 6 years and has served on the editorial boards of a number of other journals. His research focuses on the nature, development, and assessment of human cognitive abilities.

Randy G. Floyd, PhD, is Associate Professor of Psychology, Co-Director of the Child and Family Studies research area, and a member of the Institute of Intelligent Systems at the University of Memphis. Dr. Floyd is Editor of the *Journal of School Psychology* and has served on the editorial boards of several other journals. His research interests include the structure, measurement, and correlates of cognitive abilities, the technical properties of early numeracy measures, and the process of professional publication.

Preface

This book is a practical guide to the intellectual assessment of children and adolescents in the schools. Primarily intended for students of school psychology and for practicing school psychologists, it should also be useful for those counselors, teachers, administrators, and other school personnel involved in making decisions in schools based in part on the results of intelligence tests. In writing it, we have placed particular emphasis on empirically supported practices that will be useful within the response-to-intervention (RTI) model of school-psychological service delivery.

STATEMENT OF NEED

The assessment of intelligence has long been mandated by law for eligibility determination for special education and related services (e.g., intellectual disability [ID], specific learning disabilities [SLD], and intellectual giftedness). In the past, scores on standardized tests of intelligence (IQ tests) have been used in schools primarily as benchmarks against which to compare academic achievement or adaptive functioning. For example, central to most prior definitions of SLD is the concept of discrepancy in cognitive functioning. Children and youth have been identified with SLD when their academic performance or rate of skill acquisition falls substantially below what one would expect from their IQs, when that discrepancy is not attributable to certain exclusionary criteria (e.g., inadequate educational opportunity). However, the transition to the RTI model will have important implications for the use of intelligence tests in schools. To the best of our knowledge, at present no currently available textbook on intellectual assessment addresses the use of intelligence tests within an RTI context. One of our intentions in writing this book has been to fill that gap.

In addition to not addressing this need in the literature, many available books on intellectual assessment tend to be theoretically "agnostic" and present a range of theories and models of intelligence, thereby implying that readers should pick and choose any theory to meet their needs or personal predilections. Researchers in the field of intelligence, however, largely agree on the major substantive conclusions regarding the structure of cognitive abilities, the nature of individual dif-

ferences, and other major emphases of research. Thus, another of our intentions in writing this book has been to emphasize the predominant theory of intelligence in the field today—the psychometric approach. Although there are various other theories of intelligence, the psychometric approach has by far the most empirical support in the literature.

Finally, the available books on intellectual assessment tend to describe approaches for interpreting test scores that, we believe, overstate the utility of intraindividual (ipsative) analysis for differential diagnosis and intervention planning. Although research on intelligence and its assessment may lead to breakthroughs, the empirical data at the present time do not clearly substantiate the validity of disordinal models (aptitude × treatment interactions) to inform decision making, although they do provide support for ordinal models (high vs. low general cognitive ability). Therefore, this book presents a method for interpreting test scores that emphasizes the interpretation of global composite scores rather than individual subtests, with implications for ordinal models of intervention when applied to conditions of incomplete instruction (Braden & Shaw, 2009). In sum, we have written this book to address the need for an updated, evidence-based, user-friendly resource to meet these needs.

DESCRIPTION OF THE CONTENT

This book is a practical guide for school personnel working at either the primary level (i.e., elementary schools) or the secondary level (i.e., middle and high schools). In particular, it should prove useful to school psychologists, school counselors, school social workers, teachers, administrators, and other school personnel. The book has 13 chapters, many of which include tables and figures, as well as checklists and assessment forms that school personnel can easily integrate into their practices.

Chapter 1 examines the definition and nature of intelligence from the psychometric perspective, as well as some of the criticisms of this model.

Chapter 2 discusses how and why individuals differ in intelligence. It includes an explanation of research in the differential model, or quantitative behavior genetics, as well as the implications of this model for modifying intelligence.

Chapter 3 highlights ethical principles and standards most relevant to testing children and adolescents. In particular, it reviews the most recent ethical guidelines from the American Psychological Association and the National Association of School Psychologists.

Chapter 4 addresses the reasons for assessment, typical assessment processes, and potential influences on test performance that can be controlled during standardized testing or acknowledged in the interpretation of test results. It also provides practical screening tools that will promote the most accurate assessment of cognitive abilities through standardized testing.

Chapter 5 first highlights the standards guiding the selection and use of tests. It continues with a review of the most critical characteristics of tests, including norming and item scaling, as well as the reliability and validity of their scores. This information will promote selection of the best intelligence tests by practitioners, based in part on the age, ability level, and backgrounds of the clients they serve.

Chapter 6 focuses on interpretation of examinee responses and resulting scores. It targets interpreting data from qualitative and quantitative perspectives, using interpretive strategies

based on an understanding of the nature of cognitive abilities and relying on the scientific research base that illuminates empirically supported practices.

Chapter 7 provides a broad overview of the array of intelligence tests currently available. These include full-length multidimensional intelligence tests, which are the best known and most commonly used in research and practice, as well as nonverbal intelligence tests and brief and abbreviated intelligence tests.

Chapter 8 addresses two pathways by which assessment results are shared: psychological assessment reports and face-to-face contact with the parents or caregivers of the child or adolescent who has completed the assessment. It also addresses the process of sharing assessment results with supervisors during supervision meetings.

Chapter 9 highlights the evidence base describing the nature and correlates of psychometric *g*; advances in testing technology; and legal and practical constraints that warrant consideration of intelligence tests during use of the problem-solving model. It concludes that intelligence tests (1) will probably continue to play an important (though narrower) role in the schools than in the past and (2) should be included in modern problem-solving methods that seek answers to children's academic problems.

Chapter 10 describes the clinical condition known as ID and offers practical guidelines for assessing children suspected of having this condition. It addresses varying diagnostic and eligibility criteria for ID, offers recommendations for best practices in assessment, and discusses best practices in interpreting test results.

Chapter 11 discusses the use of intelligence tests in the identification of giftedness. We first review contemporary theories of giftedness. Following this, we address the definition of giftedness and issues in its identification.

Chapter 12 examines the different conceptualizations of SLD and the implications of these definitions for its identification. The use of intelligence tests in the identification of SLD has long been surrounded by controversy, which continues to this day. We address best practices in assessment that apply regardless of the SLD criteria used for identification.

Chapter 13 discusses best practices in the use of intelligence tests with children and youth from diverse backgrounds. After briefly reviewing research on test bias, we address the pros and cons of the most widely recommended best practices with this population.

Acknowledgments

We thank the many people who supported us during our writing of this book. In particular, we thank our wives and children—Theresa, Zachary, and Justin (JHK) and Carrie and Sophie (RGF)—for their enthusiasm about this book and their patience across the 3 years we devoted to developing it.

We also thank doctoral students in our school psychology programs who contributed literature reviews, retrieved articles, and helped with formatting and proofing. In particular, Ryan Farmer contributed substantially to two chapters in this book. In addition, Tera Bradley, Rachel Haley, Haley Hawkins, Sarah Irby, Jennifer Maynard, Phil Norfolk, Kate Price, Triche Roberson, Colby Taylor, and Isaac Woods deserve specific recognition. Furthermore, students enrolled in the 2010, 2011, 2012, and 2013 assessment practica at the University of Memphis were the sources of many of the examples we used, and they helped to refine our thinking about central issues and the resources we provided.

We are also appreciative of our colleagues and collaborators who provided source material for and feedback about our chapters; they include Dr. Thomas Fagan, Dr. Kevin McGrew, and Dr. Matthew Reynolds. In addition, Dr. Matthew Reynolds and Dr. Joel Schneider provided thorough and generous recommendations to us after reviewing a complete draft of this book. Finally, several professionals provided guidance to us as we developed and refined the screening tools described in Chapter 4. They include Courtney Farmer and Dr. David Damari. Overall, we benefited greatly from the support and assistance of these individuals.

We missed the opportunity to work with the late Dr. Kenneth W. Merrell throughout the development of this book, but we are grateful for his recruiting us to write it and for his support and feedback as we developed the book proposal. We would also like to express thanks to Acquisitions Editor Natalie Graham, Senior Production Editor Jeannie Tang, and Copyeditor Marie Sprayberry, all at The Guilford Press, for their guidance (and patience) as we developed and fine-tuned each chapter.

Contents

CHAPTER 1

What Is Intelligence?

A basic assumption underlying the "American dream" is that all people are created equal. Many Americans see success in life as resulting primarily from ambition, hard work, and good character, regardless of the circumstances into which one was born (e.g., gender, race/ethnicity, and socioeconomic status). This commitment to the concept of equality underlies much of the controversy that has long surrounded research and theory on intelligence and its assessment (e.g., Cronbach, 1975; Gould, 1996). Despite these well-intentioned egalitarian ideals, today's society is heavily oriented toward intelligence and intellectual achievement (e.g., Browne-Miller, 1995; Gifford, 1989). Ambition, hard work, and good character will not guarantee success in life; intelligence is also required (e.g., Gottfredson, 2011; Nisbett et al., 2012).

In schools, the assessment of intelligence is mandated by current federal special education law—the Individuals with Disabilities Education Improvement Act of 2004 (IDEA, 2004)—for the identification of intellectual disability (ID). In addition, IDEA allows, but does not require, the use of intelligence tests for the identification of specific learning disability (SLD). Finally, although gifted students are not protected under IDEA in its current form, exceptionally high intelligence (IQ) is a key component used in most states to identify intellectual giftedness (McClain & Pfeiffer, 2012). In these instances, the overall score on standardized IQ tests is primarily used as the benchmark against which to compare students' current academic achievement or adaptive behavior (i.e., age-appropriate ability to act independently, interact socially with others, care for oneself, etc.). For example, ID is identified when an individual has significantly below-average IQ and comparable deficits in adaptive functioning, among other criteria. In contrast, SLD has traditionally been identified when there is a significant discrepancy between an individual's level of academic performance and IQ, and when that discrepancy cannot be explained by certain exclusionary criteria (e.g., inadequate educational opportunities and sensory disorders). For identification of giftedness, high scores on intelligence tests are seen as necessary but not sufficient for high accomplishment. In addition to a sufficiently high level of general intelligence, other factors are required, such as motivation, creativity, and task commitment (e.g., Reis & Renzulli, 2011). Chapters 10–12 provide more detailed information on the use of intelligence tests in schools for the determination of eligibility for special education and related services.

The purpose of this chapter is to define *intelligence* and to describe how people differ in terms of their intellectual abilities. We begin by addressing two common objections to intelligence research and assessment. The first objection asserts that because intelligence is not a real thing, how can we measure something that does not exist? The second objection contends that since there is no consensus definition of intelligence, how can we measure something we cannot define?

DOES INTELLIGENCE EXIST?

The great American psychologist E. L. Thorndike (1874–1949) stated that "whatever exists at all exists in some amount" (Thorndike, 1918, p. 16, as quoted in Eysenck, 1973). Many laypersons also believe that intelligence exists as a "real thing" that underlies intelligent behavior (Berg & Sternberg, 1992). Treating an abstract concept as a concrete, physical entity, however, is a common mistake in reasoning known as *reification*. Intelligence, like gravity, is nothing more than an idea, or construct, that exists in the minds of scientists. Scientific constructs are hypothetical variables that are not directly observable.

Individuals exist, of course, and their behavior can be observed and measured. Careful examination of these measurements leads to the development of constructs that attempt to explain these factual observations. The appropriateness and usefulness of these constructs depends upon the degree to which they help us to understand, describe, and predict behavior. Thorndike, therefore, was wrong: Intelligence is not a "real thing" that exists in some amount. It is a hypothetical construct that scientists have posited to explain certain types of behavior. *Intelligence*, then, exists, but only as a scientific construct.

HOW CAN WE MEASURE SOMETHING WE CANNOT DEFINE?

Despite over 100 years of theory and research, the field of psychology has never reached a consensus on a definition of *intelligence* (e.g., see Sternberg & Kaufman, 2011). No other psychological phenomenon has proven harder to define than intelligence. In 1921, the *Journal of Educational Psychology* published a symposium titled "Intelligence and Its Measurement" (Buckingham, 1921). In this symposium the editor asked 14 leading experts to define intelligence, among other questions. He received 14 different replies. Sample responses from selected experts included the following:

Despite over 100 years of theory and research, the field of psychology has never reached a consensus definition of *intelligence*.

- "The power of good responses from the point of view of fact." (Thorndike)
- "The ability to carry on abstract thinking." (Terman)
- "Intelligence involves the capacity to acquire capacity." (Woodrow)
- "The ability of the individual to adapt himself adequately to relatively new situations in life." (Pintner)
- "Intelligence is what the tests test." (Boring)

The main source of disagreement in this symposium concerned whether intelligence is one single general ability or a number of different abilities. The only point on which the respondents tended

to agree was that intelligence is related to "higher mental processes," such as abstract reasoning or problem solving.

In 1986, Sternberg and Detterman published a book titled *What Is Intelligence?* They were interested in determining whether there was greater consensus among contemporary researchers than there had been 65 years earlier. They asked 24 prominent experts in the field of intelligence to respond to the same questions that the participants answered in the 1921 symposium. Here are a few sample definitions of intelligence from these experts:

- "Intelligence is a quality of adaptive behavior." (Anastasi)
- "Intelligence is the repertoire of knowledge and skills available to a person at a particular point in time." (Humphreys)
- "Intelligence is defined as the general factor . . . of psychological tests." (Jensen)
- "Intelligence is the reification of an entity that does not exist; it is a number of somewhat independent broad abilities." (Horn)

As Sternberg and Detterman (1986) noted, "striking diversity" was apparent among the respondents' definitions, despite over half a century of research on the nature and measurement of intelligence since the 1921 symposium. Nonetheless, consistent with the earlier findings, most of the participants mentioned that intelligence is related to higher-order cognitive functions. Differences among experts were also prevalent on the "one versus many" question of the generality of intelligence.

Conceptions of Intelligence

One reason for this lack of consensus stems from the fact that one can view intelligence from different perspectives (see Eysenck, 1998). Figure 1.1 shows the relationship among three different conceptions of intelligence. The first concerns the biological substrate that underlies all intelligent

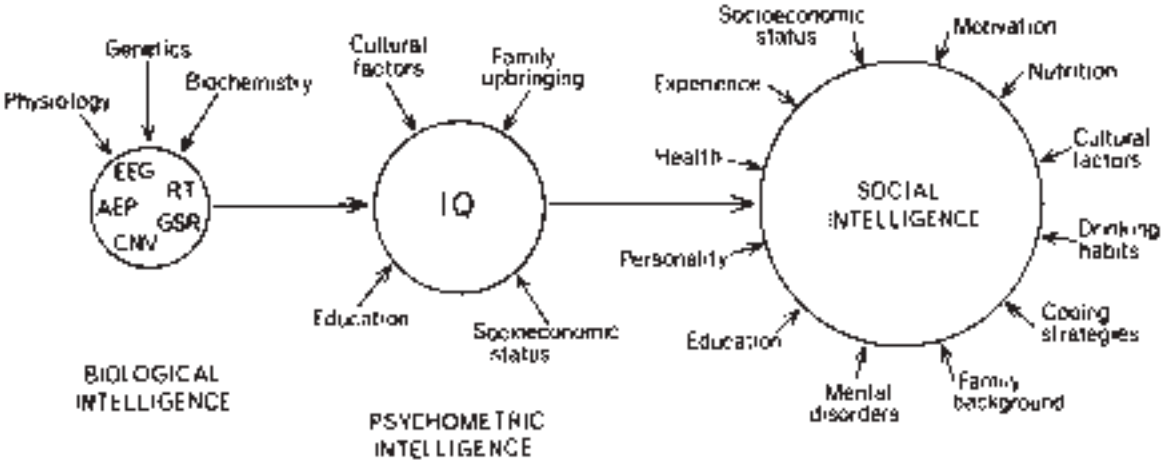


FIGURE 1.1. Relations among biological, psychometric, and social (or practical) intelligence. From Eysenck (1998, p. 62). Copyright 1998 by Transaction Publishers. Reprinted by permission.

thought and action. *Biological intelligence* sets the limits of intellectual development. Intelligence is influenced by genetics, because genes determine the neurological structures and the physiological and biochemical functioning of the brain. As shown within the circle for biological intelligence, brain functioning can be measured in a number of different ways, including the electroencephalograph (EEG), averaged evoked potentials (AEP), galvanic skin response (GSR), central nerve conduction velocity (CNV), and reaction time (RT) on elementary cognitive tasks, among other methods (e.g., see Colom & Thompson, 2011; Deary, Penke, & Johnson, 2010; Haier, 2011; Nisbett et al., 2012).

The second conception of intelligence is psychometric intelligence. *Psychometric intelligence* refers to the "intelligent" behavior that is sampled on standardized tests of intelligence ("IQ tests"). Individual differences in biological intelligence can be measured only indirectly by psychometric intelligence tests. The behaviors measured on these tests are related to biological functioning, but they are also influenced by one's background and experience. Cultural factors, family upbringing, level and quality of education, and socioeconomic status are all importantly related to intelligence test performance (e.g., Gottfredson, 2008; Jensen, 1998).

The third and final conception of intelligence is social (or practical) intelligence. *Social intelligence* refers to the overt behavior that is considered "intelligent" in specific contexts, cultures, or both. Examples include academic achievement in school and performance at work. What is considered intelligent behavior may differ to some degree across contexts and cultures, even though the basic cognitive and biological processes underlying such behavior may be the same. Intelligent behavior in the "real world" is determined in part by biological intelligence and in part by background and experience, but also by a host of other noncognitive factors (e.g., personality, healthy lifestyle, and mental health).

Thus, although these three meanings of the term *intelligence* overlap to a considerable degree, they differ in their breadth or inclusiveness. Given that there are different ways in which to view intelligence, it is perhaps not surprising that a consensus definition has eluded the field. It is important to note that this lack of consensus worries scientists much less than it appears to worry journalists, the mass media, and other laypersons. This is because scientists are more aware that consensus definitions come at the end of a line of investigation, not near the beginning when inquiry is still in the formative stages.

In any case, the key point to keep in mind is not whether tests of intelligence measure something that we all agree upon, but whether we have discovered something that is worth measuring. As we shall see in Chapter 2, despite the absence of a consensus definition of intelligence, the vast amount of research that has been conducted over the past century clearly supports its meaningfulness as a scientific construct that can be assessed with considerable accuracy. The bottom line is this: We use intelligence tests because they predict important social outcomes better than anything else that we can currently measure independently of IQ (e.g., Gottfredson, 2011).

THEORIES OF INTELLIGENCE

Scientific knowledge consists of the gradual accumulation of information by different researchers, from different types of research, and in different domains. Scientific theories are essentially "bold conjectures" that attempt to explain the known evidence in a field of study. According to philosopher of science Karl Popper (1968), "good" scientific theories make important and useful

predictions that can be subjected to empirical tests. Theories that are not potentially falsifiable are known as *pseudoscientific* theories (e.g., Freud's psychoanalytic theory). Theories of *intelligence* are scientific theories that attempt to explain differences among individuals in their ability to solve the myriad problems people confront almost daily, to learn from those experiences, and to adapt to a changing environment (e.g., Neisser et al., 1996). Theories of intelligence, however, vary in the degree to which they have been substantiated and, as a result, the extent to which they are accepted in the scientific community.

The Psychometric Paradigm

At present, the scientific theory that has by far the most empirical support is the psychometric approach to the study of intelligence. Although there are various competing theories of intelligence in the literature (e.g., see Ceci, 1990; Gardner, 1983; Das, Naglieri, & Kirby, 1994; Sternberg, 1985), the psychometric paradigm "has not only inspired the most research and attracted the most attention (up to this time) but is by far the most widely used in practical settings" (Neisser et al., 1996, p. 77).

Psychometrics is the scientific field of inquiry that is concerned with the measurement of psychological constructs. Although psychometricians study other psychological phenomena (e.g., personality, attitudes, and beliefs), the definition and measurement of intelligence have been a primary focus. According to the psychometric paradigm, intelligence involves the ability to reason abstractly, solve complex problems, and acquire new knowledge (e.g., Neisser et al., 1996). The construct of intelligence, therefore, reflects more than "book learning" or "test-taking skill" (e.g., Eysenck, 1998; Gottfredson, 2002; Neisser et al., 1996). Contrary to the view of expert opinion often reported by the mass media, most scholars with expertise in the field of intelligence agree with this working definition of intelligence (Snyderman & Rothman, 1987).

According to the psychometric paradigm, intelligence involves the ability to reason abstractly, solve complex problems, and acquire new knowledge.

Psychometric g

Two facts of nature must be explained by any theory of intelligence—and both are adequately addressed in the psychometric paradigm. The first is known as the *positive manifold* (Spearman, 1904). The positive manifold refers to the fact that all tests of cognitive ability (i.e., objectively scored tests on which individual differences are not due to sensory acuity or motor deftness) are positively intercorrelated. This means that, on average, individuals who score high on one kind of cognitive test will tend to score high on *every* other kind of cognitive test, and vice versa. The existence of the positive manifold indicates that all tests of cognitive ability share something in common that is related to differences in performance (i.e., an underlying source of individual differences, or variance).

The second fact of nature that must be explained pertains to the description of the underlying structure of these positive intercorrelations. Specifically, the central question here is whether these correlations are related to a single cognitive ability or to a number of different abilities. Charles Spearman (1904) hypothesized that the positive correlations observed among all tests of

cognitive ability result from a single underlying general ability that is shared by all tests. In other words, he asserted that all cognitive tests correlate positively because they measure the same thing to some degree. He invented factor analysis to estimate this underlying source of variance, which he called the *general factor*, or simply psychometric *g*. The purpose of factor analysis is to determine whether the correlations among a number of cognitive tests can be explained by some smaller number of inferred hypothetical constructs (i.e., factors; Carroll, 1993). Specifically, Spearman hypothesized that every test measures *g* and one other ability that is unique to that specific test (or highly similar tests).

Figure 1.2 displays Spearman's original two-factor theory. This figure shows the relationship among nine tests of cognitive ability and psychometric *g*. As can be seen, each of these tests overlaps to some degree with *g* (as expected, given the positive manifold), but not with each other. Also note that some tests are more closely related to *g* than others. S_1, for example, overlaps to a considerable extent with psychometric *g*, whereas S_2 is related to *g* to a much lesser degree. The portions of S_1 and S_2 that do not overlap with psychometric *g* reflect the portion of individual differences that are specific to each respective test.

What is psychometric *g*? Perhaps the most undisputed fact about *g* is that it is related to information-processing complexity (e.g., see Jensen, 1998). Spearman discovered that the tests correlating most highly with *g*—or the most *g*-loaded tests—are those that involve what he called *abstractness* and the *eduction of relations and correlates* (Spearman & Jones, 1950). Abstractness refers to ideas and concepts that cannot be perceived directly by the senses. The eduction of relations and correlates pertains to the perception of relations through inductive or deductive reasoning, as opposed to the simple reproduction of known rules. For example, take tests of cognitive ability that involve mathematics. Tests that correlate the most highly with psychometric *g* are those that involve solving problems, such as word problems in which the requisite arithmetic operations are not made explicit. The test taker must deduce from the description of the problem which arithmetic operations are required and then apply them. In contrast, tests that involve the

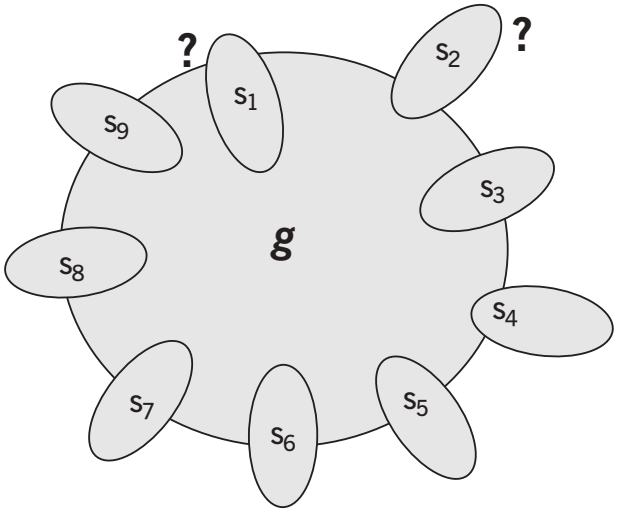


FIGURE 1.2. Spearman's two-factor theory of intelligence.

perfunctory application of explicit arithmetic operations, such as addition and subtraction worksheets, tend to be much less *g*-loaded. In other words, the relative *g*-loadedness of these kinds of tests is not related to the fact that they concern mathematics, but to the complexity of information processing required.

In addition to tasks that require inferring relations about abstract concepts, tests that correlate highly with psychometric *g* are those that require more conscious mental manipulation. A straightforward illustration of this phenomenon can be found by comparing the *g*-loadings of the Forward and Backward Digit Span tests of the Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV; Wechsler, 2003). On both tests, one must repeat a string of digits presented once at a rate of one digit per second. One repeats the digits in Forward Digit Span in the same order as presented, and in Backward Digit Span in the opposite order as presented. Thus, on Backward Digit Span, one must retain the string of digits that have been presented in short-term memory and then repeat them, just as in Forward Digit Span, but with the added requirement of reversing the digits. Because it involves more conscious mental manipulation, Backward Digit Span is more highly correlated with *g* than is Forward Digit Span (Jensen, 1998).

A test's correlation with *g* is related to the complexity of information processing required, rather than to its requisite specific skills or knowledge content.

Another noteworthy characteristic of *g* is that it cannot be described in terms of the surface characteristics of the different cognitive ability tests used in factor analysis. Spearman referred to this phenomenon as the *indifference of the indicator* (Spearman & Jones, 1950). For example, take the Block Design and Vocabulary subtests of the WISC-IV. On Block Design one must arrange colored blocks to match a pattern, and on Vocabulary one must define the meaning of different words. With their entirely different information content and response requirements, these two subtests would seem to call for quite different thought processes; yet they are two of the most highly *g*-loaded subtests on the WISC-IV (Jensen, 1998). A test's correlation with *g*, therefore, is related to the complexity of information processing required, rather than to its requisite specific skills or knowledge content.

Although it is now well established that psychometric *g* is related to processing complexity on cognitive tasks and not their surface characteristics, it is important to note that this is merely a description of *g*. Like all factors, *g* is not, strictly speaking, an explanatory construct. Factor analysis can be used to explain how various cognitive ability tests are related to each other, but it does not provide a causal explanation of the abilities or how they are organized. Therefore, the construct of *g* itself calls for an explanation. As Eysenck (1998) stated, "It is one thing to postulate a general factor of intelligence as the central conception in the theoretical framework which explains the observed phenomena in the field of mental testing; it is a different thing to postulate the nature of this factor" (p. 10).

The underlying causal mechanism of *g* is still largely unknown. Over the past 30 years, however, research at the interface between brain and behavior suggests that individual differences in intelligence are integrally related to a property of the brain known as *neural efficiency*, which is related to speed and efficiency of cognitive processing. This idea has received increased attention (e.g., Jensen, 2006; Nettelbeck, 2011). In Chapter 2, we further discuss how and why individuals differ in intelligence.

Group Factors

Not long after Spearman (1927) introduced the two-factor theory, he and other pioneers in factor analysis discovered group factors in addition to *g*. Unlike *g*, which enters into all cognitive tests, group factors are common only to certain groups of tests that require the same kinds of item content (e.g., verbal, numerical, or spatial) or cognitive processes (e.g., oral fluency, short-term memory, perceptual speed).

Thurstone's (1938) theory of *primary mental abilities* (PMAs) was the first theory of intelligence that did not include a general factor. His theory was based on the results of a new statistical technique that he developed, called *multiple factor analysis*. Multiple factor analysis initially allowed for the identification of a number of mental abilities, but no general factor. Thurstone's methodology also allowed for the identification of different kinds of cognitive abilities that were related to the factors discovered. Like Spearman's, his methodology was a product of his assumptions. Whereas Spearman's method of factor analysis was based on the assumption that only one factor is present in the correlation matrix, Thurstone's method was predicated on the notion of multiple abilities and no general ability.

Thurstone originally believed that the positive manifold is not the result of a general factor that underlies all tests of cognitive abilities, but stems from a number of fundamental abilities or PMAs. He identified seven PMAs: Verbal Comprehension, Perceptual Speed, Space Visualization, Inductive Reasoning, Deductive Reasoning, Rote Memory, and Number Facility. These factors were seen to be related to a variety of tests that shared common features. He believed that performance on any particular test, however, did not involve all the PMAs. Thurstone developed a special set of tests, called the PMA tests, to measure each of these factors.

For Thurstone's theory to be correct, each of the PMA tests had to correlate with only one factor. Furthermore, there would be no general factor with which every test correlated (representing psychometric *g*). This pattern of results in factor analysis is known as *simple structure*. Table 1.1 presents an idealized version of the results of a factor analysis showing simple structure. As can be seen here, there are six tests of cognitive ability and three factors. Each of the tests correlates perfectly with one factor and not at all with the other factors. In this case, each test represents a single group factor, uncontaminated by any other PMA. The factors are interpreted in terms of the characteristics of the tests on which they load. Thurstone's goal was to develop a battery of "factor-pure" tests such as these to measure each of the PMAs.

Results of factor analyses of Thurstone's PMA tests, however, do not contradict the existence of a general factor (e.g., see Cattell, 1971). Because all tests correlate with psychometric *g* to some

TABLE 1.1. Idealized Version of Simple Structure

	Factor		
Test	I	II	III
A	1.00	—	—
B	1.00	—	—
C	—	1.00	—
D	—	1.00	—
E	—	—	1.00
F	—	—	1.00

extent, it is only possible to approximate simple structure by allowing the factors themselves to be correlated. However, when this is done, the resulting correlations between factors can also be factor-analyzed. When these correlations are factor-analyzed, psychometric *g* emerges as a second-order factor. When factor analyses are conducted at different levels, this is known as *higher-order factor analysis*. Thus the finding that there is a psychometric *g* as well as group factors among the PMA tests indicates that Spearman and Thurstone were both correct in what they initially believed, but wrong about what they denied.

Agreement was finally reached on a general psychometric paradigm that has lasted to this day (see Carroll, 1993). According to this paradigm, individuals have a number of different abilities for solving intellectual problems and for adapting to the environment. Among these abilities, psychometric *g* is particularly important. In addition to *g*, there are specific abilities to deal with various types of problems under specific circumstances, such as those involving visual–spatial, numerical, or memory abilities. During this early period of research, a number of theories postulated by such factor-analytic luminaries as Vernon, Cattell, and Guilford attempted to describe the structure of cognitive abilities (see Carroll, 1982). Nonetheless, because opinions differed about the most appropriate method of factor analysis, consensus as to the best model of the structure of cognitive ability was not reached.

The Structure of Intelligence

One of the main reasons for this lack of consensus with regard to the structure of intelligence had to do with the fact that the early theorists were limited to a type of factor analysis called *exploratory factor analysis* (EFA). EFA is most useful in the early stages of research, when initial hypotheses about the underlying factors are generated. EFA does not test the fit between competing theories and the data within a statistical modeling framework. In the 1980s, however, use of a new statistical procedure known as *confirmatory factor analysis* (CFA) provided a way out of this theoretical dead end. CFA is a general model of factor analysis that contains all earlier models as special cases. CFA is a "sharper" factor-analytic tool than earlier models and can be used both to estimate and to test alternative factor models. In CFA, specific theories are used to construct models of the underlying structure of a battery of tests to determine how well they "fit" or explain the data. Furthermore, the "goodness-of-fit" provided by the models based on competing theories can be examined to determine which one provides the best description of the data. The results of more recent analyses using this new approach show that a hierarchical model of cognitive ability best fits the data (Keith & Reynolds, 2012; Gustafsson, 1984). In these models, both *g* and a number of group factors are represented as independent dimensions of cognitive ability (e.g., Johnson, te Nijenhuis, & Bouchard, 2008).

At the current time, the most widely accepted theory of the structure of human cognitive abilities is Carroll's (1993) three-stratum theory (e.g., Brody, 1994; Eysenck, 1994; Sternberg, 1994). Carroll's theory is based on the re-analysis of 467 data sets. It is largely an extension and expansion of the earlier theories of cognitive abilities, such as the Horn–Cattell theory of fluid and crystallized abilities (Gf-Gc; Horn, 1994; Horn & Noll, 1997). According to Carroll, "there is abundant evidence for a factor of general intelligence" (1993, p. 624), but also for a number of group factors. As shown in Figure 1.3, in the three-stratum theory psychometric *g* and group factors are arranged in a hierarchy based on their generality. *Generality* refers to the number of other factors with which a particular factor is correlated. The most general factor, psychometric *g*, is located at the

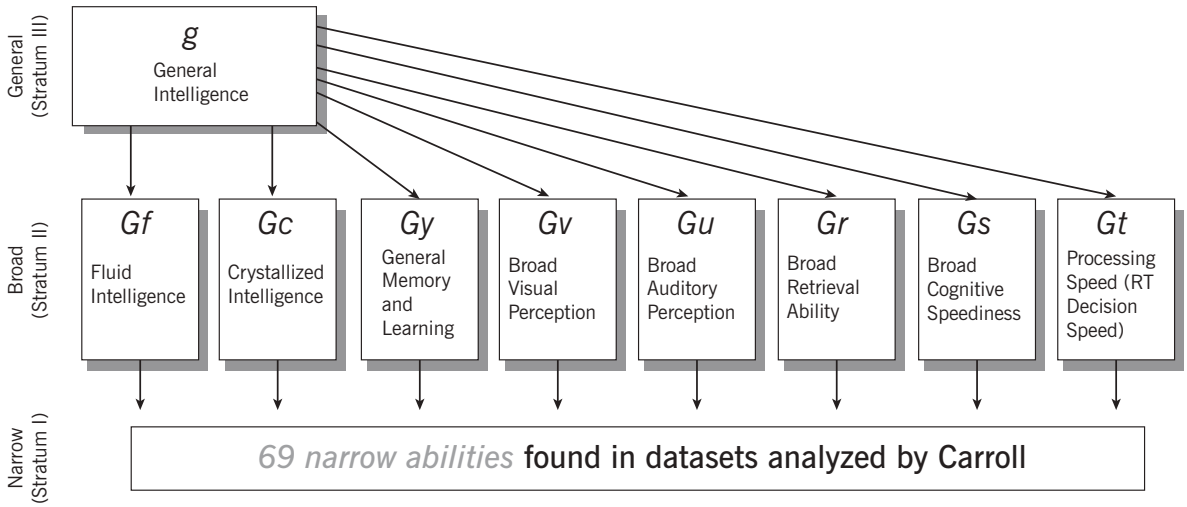


FIGURE 1.3. Carroll's three-stratum theory of the structure of human cognitive abilities. Adapted from McGrew and Flanagan (1998). Copyright 1998 by Allyn & Bacon. Reprinted by permission of Pearson Education, Inc.

top of this hierarchical structure at stratum III. Eight broad cognitive abilities (e.g., Fluid Intelligence and Crystallized Intelligence) that are similar to Thurstone's PMAs constitute stratum II; and stratum I consists of many narrow cognitive abilities (e.g., Visual Memory, Spelling Ability, and Word Fluency). *Intelligence*, therefore, is multidimensional, and consists of many different cognitive abilities (Sternberg, 1996).

In recent years, Carroll's (1993) three-stratum theory and Horn and Cattell's Gf-Gc theory have been integrated into what is referred to as the *Cattell–Horn–Carroll* (CHC) theory of cognitive abilities. This model has the same factor structure as the three-stratum theory, with a prominent psychometric *g*, but with several relatively small differences among the broad abilities posited at stratum II (e.g., Schneider & McGrew, 2012). CHC theory is increasingly being used by test developers to create theoretically driven tests that measure the general and broad factors of human intelligence.

CHC Theory and IQ Testing

How well do current intelligence tests measure the factors in the CHC theory? Given that psychometric *g* is related to the complexity of cognitive processing, all of the most widely used intelligence tests are excellent measures of *g* (see Carroll, 1993). Due to the practical limitations of psychological assessment, the battery of subtests on intelligence tests tends to be fairly small. This precludes the measurement of all factors in the CHC theory. Typically, in addition to *g*, intelligence tests measure no more than three to five broad abilities at stratum II (Keith & Reynolds, 2012). The WISC-IV, for example, measures five broad abilities (viz., Fluid Intelligence, Crystallized Intelligence, General Memory and Learning, Broad Visual Perception, and Broad Cognitive Speediness), and psychometric *g* (Keith, Fine, Taub, Reynolds, & Kranzler, 2006). Moreover, none of the intelligence tests currently in use measures all of the abilities at stratum II. At the present time, a

thorough assessment of the broad stratum II abilities requires either (1) administration of multiple intelligence tests measuring different abilities or (2) adoption of the *cross-battery approach* to intellectual assessment (e.g., Flanagan, Alfonso, & Ortiz, 2013).

What do intelligence tests measure? The global composite score on virtually every intelligence test is an excellent measure of *g*. Psychometric *g* is the largest factor on intelligence tests (e.g., Canivez & Watkins, 2010a, 2010b; Jensen, 1998; Nelson & Canivez, 2012). Psychometric *g* usually explains more variance than all group factors combined (Canivez, 2011; Kranzler & Weng, 1995), even when intelligence tests are not developed within the framework of CHC theory (e.g., Kranzler & Keith, 1999). Furthermore, the predictive validity of intelligence tests is largely a function of psychometric *g* (e.g., Gottfredson, 2002; Hunter, 1986; Jensen, 1998; Ree & Earles, 1991, 1992). On most educational and occupational criteria, the general factor usually explains from 70% to 90% of the predictable variance, depending on the criterion (e.g., Kaufman, Reynolds, Kaufman, Liu, & McGrew, 2012; Thorndike, 1985, 1986). In addition, the more complex the criterion, the more predictive psychometric *g* tends to be. Nelson and Canivez (2012), for example, found that psychometric *g* accounted for approximately 80% of the variance explained by the Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003) on a variety of academic achievement outcomes. Psychometric *g* is predictive of many important social outcomes beyond tests, such as years of education and socioeconomic status (Gottfredson, 2008). The general factor is also related (negatively) to such problem behavior as dropping out of high school, living in poverty, and incarceration (e.g., Herrnstein & Murray, 1994).

Despite the predictiveness of psychometric *g*, the predictiveness of intelligence tests is far from perfect. For example, in schools, the average correlation between measures of *g* and reading comprehension is approximately +.70. Thus the overall score on intelligence tests explains about 50% of the differences between children in the ability to comprehend text—at best. This means that up to 50% of the remaining variance must be explained by other cognitive and noncognitive variables beyond *g*, such as conscientiousness and ambition. Predictive validity coefficients between *g* and outcomes in the workplace are typically lower. As Gottfredson (1997) has stated, this implies that "the effects of intelligence—like other psychological traits—are probabilistic, not deterministic. Higher intelligence improves the *odds* of success in school and work. It is an advantage, not a guarantee. Many other things matter" (p. 116; emphasis in original).

The effects of intelligence are probabilistic, not deterministic. Higher intelligence is an advantage, not a guarantee.

Criticism of the Psychometric Paradigm

The main criticism of the psychometric paradigm is not that it is incorrect, as it has substantial empirical support, but that it is incomplete. According to Sternberg (1988a, 1988b), this approach focuses on only one (or, at most, two) of the essential aspects of intelligence, and thus cannot lead to a complete understanding of intelligence. Others, such as Ceci (1990), have argued that the syllogism implied in this line of investigation (*g* = IQ = real-world success) obscures the complex link between micro-level processing and intelligent behavior. Ceci has also contended that alternative interpretations of the data are possible. In any case, as mentioned previously, at the present time the psychometric approach—with its vast empirical support—is clearly the predominant theory of intelligence.

It is also important to note that although the three-stratum and CHC theories are now widely accepted as working models of the structure of intelligence, "the consensus is by no means unanimous, and in any case, scientific truth is not decided by plurality (or even majority) vote" (Sternberg, 1996, p. 11). While virtually every researcher within the psychometric paradigm agrees that the structure of human cognitive abilities is hierarchical, at least some question the appropriateness of an overarching psychometric *g* at stratum III of the factor hierarchy. In fact, several theories of intelligence outside the psychometric approach do not include a general ability—such as Sternberg's (1985) triarchic theory and Gardner's (1983, 1993, 1999, 2006) theory of multiple intelligences, as well as more recent theories such as the Planning, Attention, Simultaneous, and Successive (PASS) processes theory (e.g., Naglieri & Otero, 2012; Das, Naglieri, & Kirby, 1994), and Ceci's bioecological framework (e.g., Ceci, 1990). Although these contemporary theories have contributed significantly to the study of intelligence, the psychometric paradigm has been the most thoroughly researched and has by far the most empirical support.

SUMMARY

What, then, is intelligence? Most experts agree that intelligence involves the ability to reasoning and think abstractly, acquire new knowledge, and solve complex problems (Snyderman & Rothman, 1987). The vast literature on the structure of human cognitive abilities clearly indicates that it is multidimensional. This means that any single score on intelligence tests will not explain the full range of these dimensions (Sternberg, 1996). In the psychometric paradigm, however, *g* plays a particularly important role. Not only is *g* the largest factor in batteries of cognitive tests, but it is also the "active ingredient" in cognitive tests' predictiveness. Psychometric *g* is not merely linked to the surface characteristics of cognitive ability tests; rather, *g* is *directly* related to the complexity of information processing. Finally, psychometric *g* is measured quite well by most, if not all, of the most widely used intelligence tests. Certain group factors are also measured on most of these tests, but typically no more than a few are measured on any one test.

CHAPTER 2

How and Why Do People Differ in Intelligence?

It is a truism that no two people are exactly the same. Even identical twins are not *exactly* alike. Indeed, individuals differ on virtually every biological and psychological attribute that can be measured. Individual differences are observed in physical characteristics such as height, weight, blood type and pressure, visual acuity, and eye coloration; they are also observed in psychological characteristics such as personality, values, interests, and academic achievement, among others. Individuals also differ in intelligence—that is, in their general, broad, and specific cognitive abilities (e.g., see Carroll, 1993).

The scientific study of how and why people differ in terms of their psychological characteristics is known as *differential psychology* (e.g., Revelle, Wilt, & Condon, 2011). In contrast to *experimental psychology*, where differences among people are seen as a source of error to be controlled, in differential psychology those differences are the focus of research. For school psychologists and other professionals working with children and youth who have learning problems, understanding individual differences in intelligence is particularly important—not only for the identification of school-related difficulties, but also for the effective planning and evaluation of school-based interventions (e.g., Fiorello, Hale, & Snyder, 2006).

The purpose of this chapter is to explain how and why individuals differ in intelligence. We begin by discussing how individuals of the same age differ in intelligence. We then discuss developmental differences. Finally, we address why individuals differ in intelligence; we consider both the heritability and the malleability of intelligence.

THE DISTRIBUTION OF INTELLIGENCE

Around 1870, the Belgium astronomer Adolphe Quetelet (1796–1874) made a discovery about individual differences that impressed him greatly. His method was to select a physical characteristic,

such as height; obtain measurements on a large number of individuals; and then arrange the results in a frequency distribution. He found the same pattern of results over and over again, for all sorts of different things. Quetelet discovered that the distribution of these characteristics closely approximated a bell-shaped curve. The symmetrical, bell-shaped curve that results from plotting human characteristics on frequency polygons is known as the *normal* (or *Gaussian*) *distribution*. The normal curve is bell-shaped, is perfectly symmetrical, and has a certain degree of "peakedness."

Sir Francis Galton (1822–1911) was the first to discover that individual differences in intelligence closely approximate the normal distribution. Figure 2.1 shows an idealized normal distribution of intelligence test scores (IQs). As can be seen here, the average (or mean) of the distribution is 100. Most individuals have scores that are near the average, with increasingly fewer scores near the extremes of each tail. The figure shows the percentage of cases in the distribution that can be found in the different areas of the normal curve when it is divided into standard deviation units. The same proportion of scores fall within the different areas of all normal curves when standard scores are used (e.g., *z* scores, *T* scores, or deviation IQ scores). This is very important from a practical standpoint, because it allows us to determine the percentage of cases in the distribution that fall below or above any particular score, or the percentage of cases falling between any two scores.

Given that the percentage of scores in a normal distribution always equals 100%, a person's IQ can be understood in terms of its *percentile rank*. Percentile ranks are frequently used to describe standardized test scores, and for good reason: They are the easiest kind of score to understand. The percentile rank of a score is simply the percentage of the whole distribution falling below that score on the test. An individual whose score is at the 75th percentile scored higher than about 75% of the

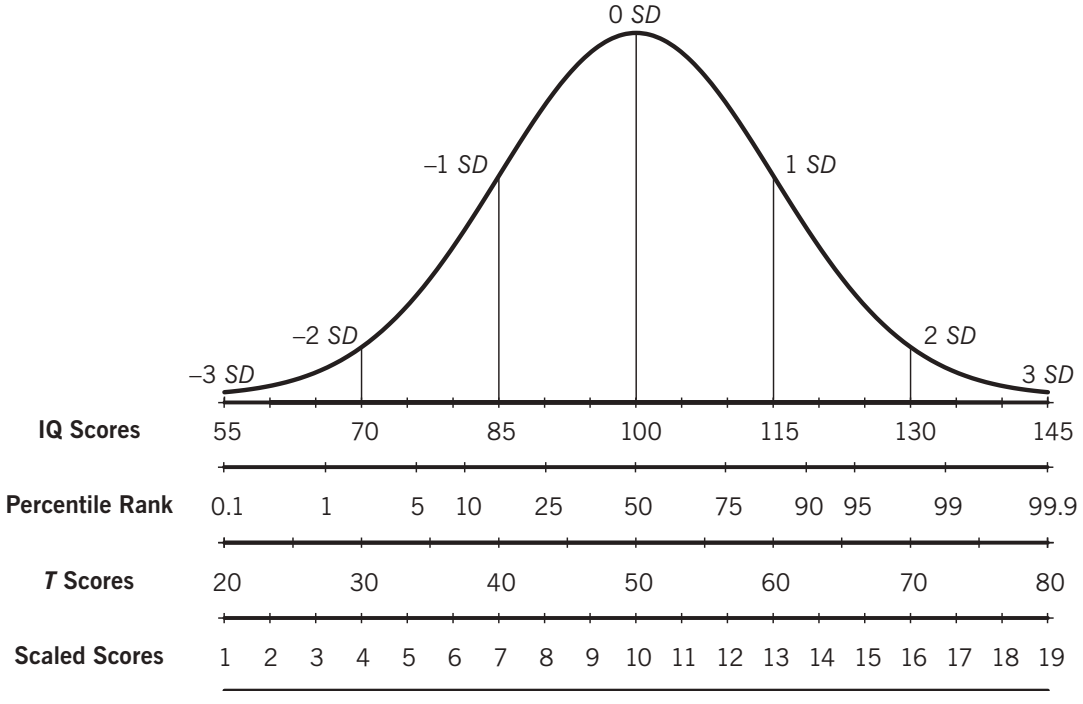


FIGURE 2.1. Theoretical normal distribution.

persons in the norm group; someone whose score is at the 50th percentile scored higher than about 50%; and so on. Most intelligence tests are developed to have a mean standard score of 100 and standard deviation of 15. Therefore, an IQ of 115, for example, has a percentile rank of 84, meaning that it surpasses 84% of all scores in the normative sample. Roughly 68% of the population has IQs that fall between 85 and 115.

The normal distribution serves as a reasonably accurate model of the frequency distribution of intelligence test scores. For individuals with IQs falling between 70 and 130, which accounts for approximately 95% of the population, the normal curve is a very good model of the distribution of scores. Nevertheless, when the 5% of scores at the extreme tails of the distribution are included, the normal distribution is not perfectly accurate. Much of the deviation from the normal curve for intelligence is due to the fact that there is a greater proportion of individuals with IQs at the very low end of the distribution (i.e., below 50 or so) than predicted by the normal curve.

The normal distribution serves as a reasonably accurate model of the frequency distribution of intelligence test scores.

Intellectual disability (ID) is typically defined by IQs below 70 or 75 (among other criteria; see Chapter 10). Individuals with mild ID (i.e., with IQs between 55 and 70) represent the lower tail of normal variability in intelligence. Severe and profound ID (corresponding to IQs below 50), however, does not tend to result from normal variability (e.g., Percy, 2007), but from either accidents of nature (i.e., single-gene defects or chromosomal abnormalities) or brain damage (e.g., birth complications, extreme nutritional deficiency, or head injury). These anomalous conditions tend to have such dramatic effects that they completely outweigh the usual genetic and environmental factors that determine a person's intelligence. Nevertheless, for most practical purposes, the normal curve provides a reasonable approximation to the distribution of intelligence—particularly when we are dealing with individuals within two standard deviations of the mean of the distribution.

DEVELOPMENTAL DIFFERENCES IN INTELLIGENCE

Alfred Binet (1857–1911) is widely regarded as the father of modern intelligence tests. The French Ministry of Education commissioned Binet and Théophile Simon (1873-1961) to find a way to identify children and youth who were at risk for educational failure. Binet and Simon began with the assumption that children become more intelligent with age, because older children are capable of doing things that younger children cannot. They developed a set of short questions and simple tasks that most children were capable of successfully completing at different ages. They then administered these items to representative samples of children at 1-year intervals from the ages of 3 to 15 years. Binet and Simon grouped items based on the percentage of children in each age group who got each item correct. For example, if a particular test item could be solved by the average child at the age of 6 years, then the age level of the problem was 6 years. Any child who passed this item was said to have a *mental age* (MA) of 6 years. They identified five items at each age level, so that they could measure MA in months as well as years. By comparing MA to *chronological age* (CA), Binet and Simon were able to determine whether a child was above or below average in relation to same-age peers. They found that at each age level, the distribution of scores on their intelligence test was approximately normal.

Wilhelm Stern (1871–1938) was the first to describe intelligence by the ratio of MA to CA to derive a *mental quotient*, which was later changed to the *intelligence quotient*, or IQ. IQ was obtained by multiplying by 100 to remove the decimal point; that is, IQ = MA/CA × 100. Thus an 8-year-old child with a MA equivalent to the average 10-year-old would have an IQ of 125 (10/8 × 100); an 8-year-old child with a MA equivalent to the average 8-year-old would have an IQ of 100 (8/8 × 100); and an 8-year-old with a MA equivalent to the average 6-year-old would have an IQ of 75 (6/8 × 100). Despite its appeal, it quickly became apparent that the new IQ scale was not useful after about 16 years of age, given that intellectual skills do not continue to develop steadily after that age. For adults this method is clearly inappropriate, because their MA would remain relatively constant, while their CA would constantly increase with time. Although many teenagers would perhaps believe in the veracity of the consistently decreasing IQ of their parents, this is clearly inappropriate. For this reason, all current intelligence tests use a point scale in which an individual's score is compared to a normative sample of same-age peers.

Figure 2.2 shows the theoretical growth curve of intelligence, based on repeated testing of the same individuals over time. As shown in the figure, between the ages of 4 and 12 years or so, the rate of cognitive growth is fairly linear. After that age, the rate of gain begins to decrease gradually as the individual reaches maturity, between 16 and 20 years of age. The development of intelligence, therefore, is similar to that of height. Although scores on IQ tests are relatively unstable during infancy and the preschool years, they stabilize rather quickly during the early school years and remain that way throughout adulthood (e.g., Humphreys, 1989). This means that the rank order of intelligence test scores will tend to remain relatively the same as individuals move from childhood to adulthood. Prior to age 4 or 5, however, the long-term stability and predictive power of intelligence tests are of less value (Goodman, 1990). In addition, the stability of test scores decreases with time. More recent test results will be a better indicator of a person's current cognitive ability than tests administered in the distant past.

Figure 2.3 shows the growth curves for fluid general intelligence for three hypothetical individuals. Although the cognitive ability of all three increases over time, plateauing at about the age of 15 years, the relative ranking of each person remains the same. Because children's IQs are

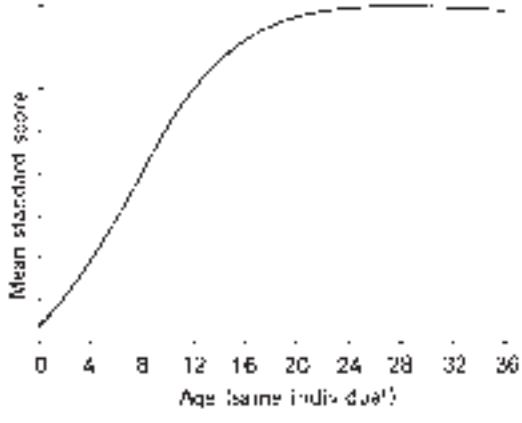


FIGURE 2.2. Theoretical growth curve based on repeated testing of the same individual. From Eysenck (1979, p. 59). Copyright 1979 by Springer-Verlag. Reprinted by permission.

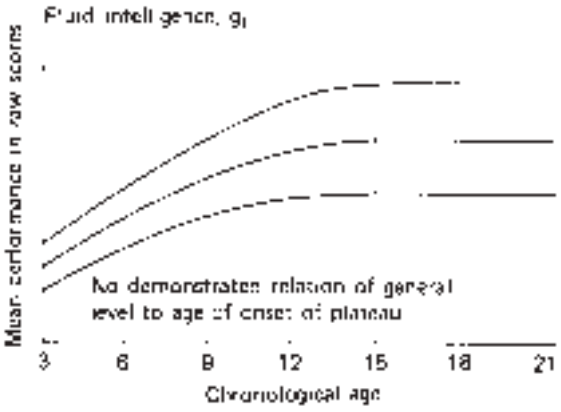


FIGURE 2.3. Growth curve of fluid ability (Gf) with age. From Eysenck (1979, p. 75). Copyright 1979 by Springer-Verlag. Reprinted by permission.

derived from comparison to same-age peers, this means that children's scores on intelligence tests will remain relatively the same, despite the fact that their cognitive ability is steadily developing. An 8-year-old and a 16-year-old with IQs of 100 are not the same in terms of their general reasoning abilities. On the contrary, it simply means that they are average for their respective age groups. Students who obtain the same IQ from one age to another have simply maintained their relative standing in relation to same-age peers, while their cognitive abilities may have developed considerably. It is also important to note that although IQ scores are generally quite stable over time, the consistency of scores is not perfect. Just like height, many individuals' level of intelligence may remain roughly constant in relation to that of same-age peers over time, but others' may not. The average change in IQs for individuals between 12 and 17 years of age, for example, is about 7 points, but for some people it can be as much as 18 points (e.g., Neisser et al., 1996).

Although intelligence test scores are generally quite stable over time, the consistency of scores is not perfect.

Although much is known about the development of intelligence through childhood, research on change in cognitive abilities in adulthood and old age is less clear—largely due to methodological issues surrounding the use of cross-sectional research designs, cohort effects, and which cognitive abilities are to be measured and how (e.g., see Salthouse, 2010). Nevertheless, examination of the age norms for widely used intelligence tests among adults, such as the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV; Wechsler, 2008), are fairly straightforward (Deary, 2001). For some subtests on the WAIS-IV, there is little change across age among adults; for other subtests, age-related decrements in ability are apparent. The subtests on which older and younger people tend to compare well are vocabulary, general information, and verbal reasoning. These subtests draw upon one's knowledge base that has been accumulated through experience over time (i.e., Crystallized Intelligence, or Gc).

The cognitive subtests on which marked differences between older and younger people are observed are those that tend to be speeded or involve abstract or spatial reasoning. These subtests measure current cognitive functioning, because they involve solving problems here and now, often with novel content and under speeded conditions (i.e., Fluid Intelligence or Gf). For example,

results from Schaie's (1996) Seattle Longitudinal Study for verbal ability (Gc) and inductive reasoning indicated that average scores on verbal ability remained fairly constant across age. On inductive reasoning, in contrast, average scores decreased considerably across age. It is important to note that these were average changes in inducting reasoning. For some people, performance on Gf tasks such as inductive reasoning does not change or improves with age. What does age affect? One theory proposed by Salthouse (1996) is that age-related changes in cognitive abilities are caused by slowing in the speed of information processing in the brain. Salthouse found that the effects of age on the brain were primarily on the general factor of intelligence (psychometric *g*). Changes observed in the measures of Gf and Gc across the lifespan are due to their relationship to *g*, not something specific to each broad ability at stratum II of the three-stratum theory (see Carroll, 1993, and Chapter 1, this volume). Further research on the study of cognitive aging is needed, however—particularly given that life expectancy is increasing dramatically, and thus the proportion of older people in the population is growing.

WHY DO PEOPLE DIFFER IN INTELLIGENCE?

Given the controversy that has surrounded research on the so-called "nature–nurture" issue in intelligence over the years, it may come as a surprise to learn that the main focus of research in this area no longer involves the question of whether individual differences in intelligence are related to genes or the environment (Nisbett et al., 2012). Results of surveys indicate that most psychologists believe that *both* genetic and environmental factors are at least partly related to variability in intelligence (Snyderman & Rothman, 1987). As Ceci (1990) has put it, "there are so many reports in the behavior-genetics literature of substantial heritability coefficients, calculated from an enormous variety of sources, that it is unlikely they could all be wrong" (p. 130). Rather, the focus of much contemporary research is on the identification of specific genes responsible for the heritability of intelligence, and on specific pathways between genes and behavior (e.g., Plomin & Spinath, 2004). Before we discuss research on the heritability of intelligence, it is important to describe how researchers investigate the relative contributions of nature and nurture to individual differences in behavioral traits such as intelligence.

Most psychologists believe that both genetic and environmental factors are at least partly related to variability in intelligence.

Quantitative Behavior Genetics

Quantitative behavior genetics, or the *differential model*, is used to examine the contributions of genetics and the environment to individual differences in behavioral traits. Research in behavior genetics is best understood as an attempt to partition *variance*. Variance is the average of the squared deviations of every score from the mean of the distribution:

$$SD^2 = \frac{\sum \left(X - \overline{X}\right)^2}{N}$$

where Σ is "the sum of," X refers to each obtained score, $\overline{X}$ is the mean of X, and N refers to the total number of scores. The variance is interpreted as the "amount of information" in a distribution (Cronbach, 1990).

In quantitative behavior genetics, variance in an observable, measurable characteristic of an individual—the phenotype *(VP)*—is partitioned into components of variance related to the genotype *(VG)* and the environment *(VE)*. Thus $VP = VG + VE$. The proportion of variance in the phenotype that is attributable to variance in the genotype is known as *heritability* (h^2) of a particular trait ($h^2 = VG/VP$). *Environmentality* reflects differences in the phenotype that are associated with the environment and with measurement error (i.e., $1 - h^2$). The environmental component of variance can be further partitioned into two subcomponents, *shared* and *nonshared*. Shared environmental influences reflect common experiences that make individuals in the same family similar to each other and different from those in other families (e.g., socioeconomic status, parenting style). Nonshared environmental influences reflect unique life experiences that make members of the same family different from one another (e.g., peer groups). In addition, phenotypic variance may also be influenced by combined genetic and environmental influences, such as genotype–environment correlations. Genotype–environment correlations occur when people's genetic propensities are related to the environments they experience. A common example in the schools is the identification and placement of children in special classes for the intellectually gifted.

All procedures for estimating the relative contributions of heredity and environment in quantitative behavior genetics involve correlating measurements taken from groups of people who are and are not biologically related, and then comparing these correlations with those expected from a purely genetic hypothesis. The genetic correlation for identical, or *monozygotic* (MZ), twins is 1.00, because they share 100% of their genetic makeup. Since non-twin offspring (i.e., full siblings) receive 50% of their genes from each parent, the parent–child genetic correlation is .50. The genetic correlation for fraternal, or *dizygotic* (DZ), twins is also .50, because they too share 50% of their genes. Persons who are not biologically related have no genes in common, so the genetic correlation between these individuals is .00. When the correlations for a particular kinship (e.g., MZ twins) differ substantially from the predicted genetic correlation, results do not substantiate a purely genetic hypothesis. If there is no resemblance between family members, then the genetic hypothesis is disconfirmed for that particular trait.

Family, twin, and adoption studies are the three basic research designs used in the differential model. These designs essentially exploit different combinations of genotypes and environments that occur naturally in the population. In family studies, estimates of h^2 are based on comparisons of parents with their offspring and of siblings with each other. Because members of the same family share both common genes and environment, evidence of familial similarity is consistent with the genetic hypothesis, but cannot confirm it. Twin and adoption studies must be used to separate the relative contributions of genetics and the environment to phenotypic variance, because they essentially control for the effects of genes or the environment.

Twin studies are one of the most powerful designs in the differential model. If a particular trait is influenced by genetics, then the correlation between MZ twins should be greater than the correlation between DZ twins, because MZ twins share 100% of their genes and DZ twins share only 50%. Correlations for MZ and DZ twins reared together reflect the phenotypic variance accounted for by the combined effects of genes and common environment. Adoption studies are another invaluable source of information on the relative contributions of genetics and the

environment to phenotypic variance, because they too separate the effects of common genes and common environment. In adoption studies, adoptive parents are compared to their adopted children, to whom they are completely genetically dissimilar, and biological parents are compared to their adopted-away children, with whom they have no environmental experiences in common. Any resemblance between adopted children and their adoptive parents, therefore, can be attributed to the effects of shared environment; conversely, any resemblance between adopted-away children and their biological parents can be attributed to genetics. Pairs of adopted and unadopted children reared together can also be compared in adoption studies to estimate the effects of common environment. Particularly noteworthy are studies in which the adoption and twin methods are combined. In these studies, of which there are relatively few, MZ and DZ twins who have been separated at birth and reared apart are compared.

Heritability of Intelligence across the Lifespan

Bouchard and McGue (1981) reviewed the literature of familial studies of the heritability of intelligence. They reviewed over 100 independent studies, summarizing over 500 correlations among more than 100,000 family members. Taken as a whole, results of these studies were consistent with a genetic hypothesis, but provided support for the influence of the environment on individual differences on intelligence (see Table 2.1). Not only were the correlations between MZ twins greater than those between family members with less genetic overlap, even when MZ twins were reared apart, but the obtained correlations across kinships did not deviate substantially from the expected genetic correlations. Evidence for the role of environment was reflected in the fact that correla-

TABLE 2.1. Average IQ Correlations among Family Members

Relationship	Average *r*	Number of pairs
	Reared-together biological relatives	
MZ twins	.86	4,671
DZ twins	.60	5,533
Siblings	.47	26,473
Parent–offspring	.42	8,433
Half-siblings	.35	200
Cousins	.15	1,176
	Reared-apart biological relatives	
MZ twins	.72	65
Siblings	.24	203
Parent–offspring	.24	720
	Reared-together nonbiological relatives	
Siblings	.32	714
Parent–offspring	.24	720

Note. MZ, monozygotic; DZ, dizygotic. Adapted from McGue, Bouchard, Iacono, and Lykken (1993). Copyright 1993 by the American Psychological Association. Adapted by permission.

tions between biological relatives reared together were greater than those between individuals of the same degree of kinship reared apart; and correlations between biologically unrelated persons reared together were substantially greater than zero. Neither of these findings can be explained by genetics alone. For these data, h^2 for general intelligence was estimated to be .51 (Chipeur, Rovine, & Plomin, 1990), indicating that slightly more than half the phenotypic variance was accounted for by genetic factors. Shared environmental influences accounted for 11–35% of the variance, depending on kinship data used in the estimation. Nonshared environmental variance was found to explain 14–38% of the variance. Therefore, the results of this review indicated that environment and genetics contribute roughly equally to variability in psychometric *g*.

A later review of the literature by McGue, Bouchard, Iacono, and Lykken (1993), however, reached a somewhat different conclusion regarding the relative contributions of heritability and environmentality to individual differences in general intelligence across the lifespan. They found that the correlation between the IQs of MZ twins increased across the lifespan. The correlation between DZ twins, in contrast, although fairly stable between 4 and 20 years of age, decreased dramatically after late adolescence. Interestingly, subsequent research has shown that shared environmental influences account for between 30% and 40% of the variance in IQ until the age of about 20 years, after which the amount of variance explained drops to zero (Loehlin, 2007). Nonshared environmental effects, in comparison, remain fairly constant across the lifespan, explaining somewhere between 10% and 20% of the variance from 4–6 years of age into adulthood.

Further substantiation that the underlying genetic and environmental determinants of IQ variability change over the lifespan comes from adoption studies. Among persons who were not biologically related, but reared together, research indicates that average correlation between siblings during childhood is rather substantial, accounting for approximately 25% of the variance in IQ (e.g., Bouchard, 2009; Nisbett et al., 2012). During adulthood, however, this correlation again decreases to zero. Taken together, these results indicate that the environmental effects influencing variability in intelligence are *nonshared*. Neisser et al. (1996) have stated:

> No one doubts that normal child development requires a certain minimum level of responsible care. Severely deprived, neglectful, or abusive environments must have negative effects on a great many aspects—including intellectual—of development. Beyond that minimum, however, the role of family experience is in serious dispute. (p. 88)

Interpreting Heritability Coefficients

Several important points must be kept in mind when we are interpreting h^2 estimates, however. First, h^2 is a statistic that describes the ratio of genotypic variance to phenotypic variance in a population. Although it does indicate that physiological functioning is related to individual differences in the trait, it provides no information on the specific areas of the brain responsible for that variation or how they operate. Second, estimates of h^2 refer to populations, not to individuals. An h^2 of .50 does not indicate that half of an individual's measured trait is attributable to genes and half to the environment. Third, estimates of h^2 that are greater than zero do not imply biological determinism. Genotypes do not directly cause phenotypic expression. As Rushton (1995) has stated, "they code for enzymes, which, under the influence of the environment, lay down tracts in the brains and nervous systems of individuals, thus differentially affecting people's minds and the choices they make about behavioral alternatives" (p. 61).

Malleability of Intelligence

The effects that genes have on individual differences in intelligence are better described as genetic influence than as genetic determinism. Being born with a particular set of genes does *not* predetermine behavioral traits. Particular genotypes do not correlate perfectly with expressed phenotypes. This is clearly seen in the correlations for identical twins reared together that are less than perfect. Despite the fact that identical twins share the same genotype, their expressed level of intelligence is highly correlated, but not exactly the same. These differences can only be explained by the environment. "Gene action always involves an environment—at least a biochemical environment, and often an ecological one" (Neisser et al., 1996, p. 84). Thus, traits with substantial heritable components, such as intelligence, are susceptible to environmental intervention.

Behavior geneticists use the concept of *reaction range* (RR) to explain variability in phenotypes for a particular genotype. Figure 2.4 shows the hypothetical effects of three different environments (viz., restricted, natural, and enriched) upon four different genotypes for the expression of intelligence. For all genotypes, the level of intelligence that is expressed is lower when individuals are raised in a deprived environment and higher when they are brought up in an enriched environment. The RR for any genotype is limited, however, as is reflected in the vertical lines on the right of the figure. As can be seen here, within any particular environment there will be differences among individuals, depending upon their genotypes. Also, the RR is larger for some genotypes than for others. More specifically, there is a wider RR for those with more advantaged genotypes. Genes, therefore, do not determine behavior; instead, they determine a range of responses to the environment.

Over the course of development, environmental effects can give rise to considerable variation in the phenotypic expression of any single genotype. Nevertheless, as Jensen (1981) has asserted, "there are probabilistic limits to the reaction range, and one of the tasks of genetic analysis is to explore the extent of those limits in the natural environment and to discover the environmental

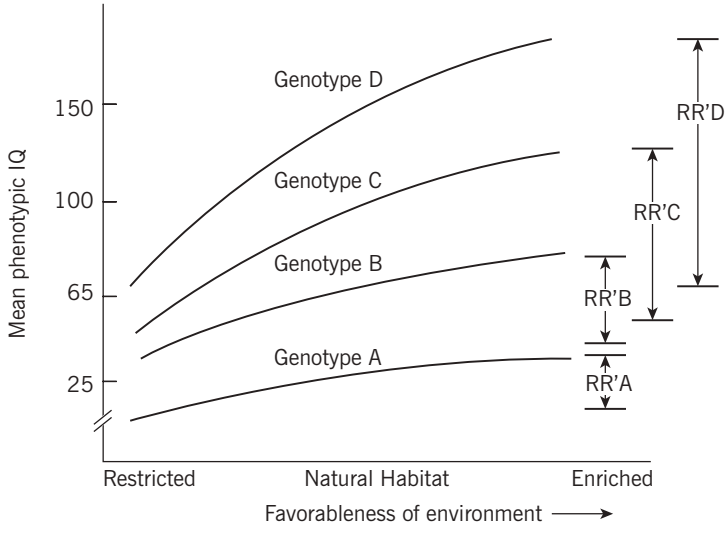


FIGURE 2.4. Hypothetical reaction range (RR) for four genotypes. Adapted from Platt and Sanislow (1988). Copyright 1988 by the American Psychological Association. Adapted by permission.

agents that affect them" (p. 485). At the current time, despite numerous efforts to discover ways to increase children's intelligence, the empirical evidence indicates that there are indeed limits to the malleability of IQ (e.g., Spitz, 1999). Results of early intervention programs for preschoolers, such as Project Head Start, have consistently shown that there is a "fade-out" of any gains in IQ points within 3 years of program termination (for reviews, see Clarke & Clarke, 1989; Locurto, 1991; Spitz, 1986).

In addition, those intervention programs that have resulted in substantial and durable gains in IQs often fail to evince comparable improvement on tasks that are highly correlated with IQ. On the Milwaukee Project (e.g., Garber, 1988), for example, large differences of about 30 IQ points were observed between the experimental and control groups, but no significant differences were observed on measures of academic achievement that are closely related to intelligence, such as reading comprehension. These results suggest that these gains in IQ resulted from "teaching to the test" and did not reflect real increases in IQ (e.g., Jensen, 1989a). Results of adoption studies—which are arguably the most intensive "interventions" imaginable—suggest that gains or losses of about 10–12 IQ points are the most that can be expected from dramatic environmental interventions (e.g., Locurto, 1990, 1991; Nisbett et al., 2012; Rowe, 1994). Last, although certain cognitive skills definitely can be taught "at least to some of the people, some of the time" (Sternberg, 1996, p. 13), research has shown that such learning tends to be quite context-bound and does not transfer to other cognitive domains (e.g., Detterman & Sternberg, 1993).

Taken as a whole, research indicates that IQ is *imperfectly malleable*. With the exception of instances of extreme social isolation and neglect, most environments are functionally equivalent for the development of intelligence. Of course, the failure of most intervention programs over the past 50 years to dramatically increase IQ does not preclude novel interventions from doing so in the future. Nonetheless, "if environmental interventions are to succeed, they must be truly novel ones, representing kinds of treatments that will be new to most populations" (Rowe, 1994, p. 223).

Research indicates that IQ is *imperfectly malleable*. Most environments are functionally equivalent for the development of intelligence.

SUMMARY

Much of the research and theory on individual differences in intelligence can be summed up as follows: (1) The distribution of general intelligence is approximately normal among people of the same age; (2) intelligence develops from childhood to maturity, plateauing at about 16 years of age; (3) changes in cognitive abilities occur across the lifespan, with cognitive abilities related to Gc being the least affected and those related to Gf being the most affected; (4) both genetic and environmental factors are importantly related to individual differences in intelligence, but the relative contribution of each changes over the lifespan; and (5) IQ is somewhat, but not completely, malleable.

CHAPTER 3

Ethics in Assessment

It is important that students in training and professionals engaged in the practice of psychology read and frequently review the most recent ethical guidelines from the American Psychological Association (APA; go to *www.apa.org/ethics/code/index.aspx*). For those engaged in the practice of school psychology (especially in school settings), we also recommend referencing similar guidelines from the National Association of School Psychologists (NASP; go to *www.nasponline.org/standards/2010standards.aspx*). This chapter highlights the ethical principles and standards from these two professional groups that are most relevant to testing children and adolescents. In many ways, this chapter forms one segment of the foundation for the remainder of the chapters in this book that focus on professional practices. For broader coverage of ethical standards and principles, as well as coverage of laws and court decisions relevant to testing children and adolescents, please consider the following texts: Jacob, Decker, and Hartshorne (2011) and Sattler (2008, Chapter 3). Content from the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 1999) is discussed in Chapter 5.

THE AMERICAN PSYCHOLOGICAL ASSOCIATION

In its Ethical Principles of Psychologists and Code of Conduct, including the 2010 Amendments, the APA (2002, 2010a) addresses five general principles that should serve as aspirations for professional practice: beneficence and nonmaleficence; fidelity and responsibility; integrity; justice; and respect for people's rights and dignity. Table 3.1 presents descriptions of these general principles. The APA's Ethical Principles of Psychologists and Code of Conduct also includes 10 specific ethical standards. Table 3.2 lists each of these standards and highlights some components that are most relevant to the practice of testing. In particular, Standard 9 comprises 11 components related to assessment. These components address four themes: high-quality assessments; control of information; interpretation of assessment results; and competence.

TABLE 3.1. General Ethical Principles from the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct, and Their Descriptions

Beneficence and Nonmaleficence

Psychologists strive to benefit those with whom they work and take care to do no harm. . . .

Fidelity and Responsibility

Psychologists . . . uphold professional standards of conduct, clarify their professional roles and obligations, accept appropriate responsibility for their behavior, and seek to manage conflicts of interest that could lead to exploitation or harm. . . .

Integrity

Psychologists seek to promote accuracy, honesty, and truthfulness in the science, teaching, and practice of psychology. . . .

Justice

Psychologists . . . take precautions to ensure that their potential biases, the boundaries of their competence, and the limitations of their expertise do not lead to or condone unjust practices.

Respect for People's Rights and Dignity

Psychologists respect the dignity and worth of all people, and the rights of individuals to privacy, confidentiality, and self-determination. . . . Psychologists are aware of and respect cultural, individual, and role differences, including those based on age, gender, gender identity, race, ethnicity, culture, national origin, religion, sexual orientation, disability, language, and socioeconomic status and consider these factors when working with members of such groups. Psychologists try to eliminate the effect on their work of biases based on those factors . . .

Note.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060–1073.

American Psychological Association. (2010a). Amendments to the 2002 ethical principles of psychologists and code of conduct. *American Psychologist*, *65*, 493.

High-quality assessments are supported by three components of Standard 9. Component 9.01 states that psychologists should base their conclusions and recommendations offered in oral and written reports on evidence from the most appropriate methods of assessment. Opinions about a person's psychosocial functioning should be based on the results of a formal assessment, not on hearsay, rumors, or other sources of unverifiable information. In the same vein, Component 9.02 conveys that psychologists should rely on sound evidence when selecting which assessment techniques to use, modify, and interpret. More specifically, they should consider reliability and validity evidence (see Chapter 5) relevant to their assessment techniques and the clients they serve, and when this evidence is not strong or plentiful, they should articulate the limitations of their assessment results. Thus psychologists should work diligently in selecting,

Psychologists should consider reliability and validity evidence relevant to their assessment techniques and the clients they serve, and when this evidence is not strong or plentiful, they should articulate the limitations of their assessment results.

TABLE 3.2. Ethical Standards from the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct (Including the 2010 Amendments)

Standard 1: Resolving Ethical Issues

Standard 2: Competence

2.01 Boundaries of Competence

(b) Where scientific or professional knowledge in the discipline of psychology establishes that an understanding of factors associated with age, gender, gender identity, race, ethnicity, culture, national origin, religion, sexual orientation, disability, language, or socioeconomic status is essential for effective implementation of their services or research, psychologists have or obtain the training, experience, consultation, or supervision necessary to ensure the competence of their services, or they make appropriate referrals . . .

2.03 Maintaining Competence

Psychologists undertake ongoing efforts to develop and maintain their competence.

2.04 Bases for Scientific and Professional Judgments

Psychologists' work is based upon established scientific and professional knowledge of the discipline.

Standard 3: Human Relations

3.09 Cooperation with Other Professionals

When indicated and professionally appropriate, psychologists cooperate with other professionals in order to serve their clients/patients effectively and appropriately.

3.10 Informed Consent

(a) When psychologists conduct research or provide assessment, they obtain the informed consent of the individual or individuals using language that is reasonably understandable to that person or persons [in most circumstances].

(b) For persons who are legally incapable of giving informed consent, psychologists nevertheless (1) provide an appropriate explanation, (2) seek the individual's assent, (3) consider such persons' preferences and best interests, and (4) obtain appropriate permission from a legally authorized person, if such substitute consent is permitted or required by law.

Standard 4: Privacy and Confidentiality

4.01 Maintaining Confidentiality

Psychologists have a primary obligation and take reasonable precautions to protect confidential information obtained through or stored in any medium . . .

4.02 Discussing the Limits of Confidentiality

(a) Psychologists discuss with persons and organizations with whom they establish a scientific or professional relationship (1) the relevant limits of confidentiality and (2) the foreseeable uses of the information generated through their psychological activities.

Standard 5: Advertising and Other Public Statements

Standard 6: Record Keeping and Fees

6.01 Documentation of Professional and Scientific Work and Maintenance of Records

Psychologists create, and to the extent the records are under their control, maintain, disseminate, store, retain, and dispose of records and data relating to their professional and scientific work . . .

(continued)

TABLE 3.2. *(continued)*

Standard 7: Education and Training

Standard 8: Research and Publication

Standard 9: Assessment

Standard 10: Therapy

administering, scoring, or interpreting the results of their assessment instruments, to ensure that they facilitate the most accurate assessments and produce the best outcomes for their clients. Finally, in Component 9.08, psychologists are implored to avoid old and obsolete tests as a basis for current assessments, decisions about interventions, or recommendations. In reference to intelligence testing, tests normed more than 10 years ago should be avoided, if possible (see Chapter 5). We also assert that there is about a 2-year window in which psychologists should make the transition from an older version of a test to a more recently published version of that test, but we recognize that professional judgment must be applied in considering this transition (see Dombrowski, 2003; Lichtenstein, 2010; Oakland, 2003). Psychologists, however, must clearly stay up to date in their assessment knowledge.

Control of information is addressed in three components. Two focus on informed consent and confidentiality. Component 9.03 states that psychologists should ensure both (1) that their clients are informed about the nature and purpose of the assessment, what other parties (if any) are privy to the information yielded by the assessment, and confidentiality and its limits; and (2) that they (or, in the case of minors, their parents) consent to the assessment. Psychologists always inform clients about the kind of assessment they are about to conduct, and tell them why they are assessing them, in a language they can understand. Psychologists also obtain informed consent before conducting assessments, except when (1) the assessment is required by law or government regulations; (2) informed consent is implied as part of a customary educational, institutional, or organizational activity; or (3) one purpose of the testing is to evaluate a client's ability to make decisions independently of a guardian.

Component 9.04 targets release of assessment results and conveys that psychologists release these results to only their clients, others to whom their clients have released the results, and others as required by law or court order. Assessment results include the following: an examinee's raw and standardized scores; the examinee's actual responses to test questions and stimuli; and the psychologist's notations and records about the examinee's statements and behavior during the assessment. In addition, in Component 9.11 psychologists are implored to maintain *test security* by avoiding release of test manuals, test items, test pages, other testing materials, and test protocols,

consistent with law and contractual obligations. Because well-developed assessment instruments are the products of years of effort to develop and refine them, it is important for test users to honor these efforts. More specifically, users are prohibited not only from duplicating these resources in hard copy and digital forms without permission, but also from sharing specific test content—and even response strategies—that might inflate test scores due to invalidity in assessment (see Chapter 4). Essentially, revealing such information spoils the test's ability to differentiate those who possess the underlying knowledge or skills from those who do not; in other words, validity is undermined. Consistent with this theme, in practice administrations during training, answers to items should not be provided upon an examinee's request, and at no point should others not in training be allowed to peruse the testing materials or be advised regarding how to improve performance on such tests. It is important to note that completed test protocols are "educational records" in school settings. So, despite psychologists' being held to high standards for maintaining privacy and confidentiality of records as well as intellectual property rights, copyright interests, and test security standards, parents' rights to examine their children's records may supersede them. Thus, they may examine test protocols, including specific items, but this review is typically conducted with guidance by a competent professional. Parents, however, do not have a legal right to review the personal notes of psychologists.

Psychologists are implored to maintain *test security* by avoiding release of test manuals, test items, test pages, other testing materials, and test protocols.

Interpretation of results is addressed in three components. Component 9.06 implores psychologists to consider any attribute of the person being assessed that might undermine the accuracy of test interpretations. These attributes include external and internal influences that are not relevant to the construct targeted by the assessment, such as the client's test-taking abilities and any situational, personal, linguistic, and cultural differences (see Chapters 4 and 13). Component 9.09 reminds psychologists that they are responsible for drawing valid conclusions from the assessment instruments that they employ as part of their assessments, regardless of who administers and scores these instruments and in cases of electronic administration and scoring. In addition, Component 9.10 implores psychologists to convey assessment results clearly to clients.

Finally, competence is addressed not only in Standard 2 in general (see Table 3.2), but also in Component 9.07 of Standard 9. In it, psychologists are encouraged to ensure that psychological techniques are employed by only qualified individuals or, in the case of training, under the supervision of those who are qualified.

THE NATIONAL ASSOCIATION OF SCHOOL PSYCHOLOGISTS

In its Principles for Professional Ethics, the NASP (2010a) addresses many of the same themes, principles, and specific ethical standards as addressed in the APA's (2002, 2010a) Ethical Principles of Psychologists and Code of Conduct, but it uses slightly different terms to describe them. In this section of the chapter, the NASP's broad ethical themes I, III, and IV are discussed first, and then emphasis is placed on the broad ethical theme II because of its focus on assessment. Again, full access to the NASP's Principles for Professional Ethics (2010a) is available at *www.nasponline.org/standards/2010standards.aspx*.

General Principles

Theme I addresses three principles related to respecting the dignity and rights of all individuals to whom school psychologists provide services. These principles are most similar to the five general ones offered by the APA (see Table 3.1). They implore school psychologists to honor the rights of others to participate and to decline participation in the services they provide. For example, school psychologists should ensure that parents provide informed consent (typically, written consent) for the provision of extensive or ongoing psychological or special education services (including assessment) for their children before these services begin. Although administration of all intelligence tests (and other psychological tests) requires consent, because the tests could be seen as intrusions on privacy beyond what might be expected in the course of ordinary school activities, other assessment practices (such as review of records, classroom observations, academic screening, and progress monitoring) do not typically require parental consent, if those conducting them are employees of the district in which they practice. School psychologists should also consider seeking and obtaining children's assent (i.e., their affirmative agreement) to participate in assessment or intervention activities; such assent is not required legally or ethically, but it is best practice (Jacob et al., 2010). Furthermore, these principles guarantee the rights to privacy and confidentiality to those who participate in school psychology services, while promoting information about the limits of confidentiality. Finally, school psychologists should promote fairness and justice for all.

School psychologists should collaborate and engage in respective interactions with other professionals to promote the well-being of students and families with whom they work.

Theme III addresses honesty and integrity in interactions with parents, children, and other professionals. Its principles urge school psychologists to accurately present their competencies and to clearly convey the nature and scope of their services. School psychologists should collaborate and engage in respectful interactions with other professionals to promote the well-being of students and families with whom they work. In addition, they should strive to avoid relationships in which they may lose objectivity and diminish their effectiveness because of prior and ongoing personal relationships or their service in multiple professional roles (as evidenced through financial or other conflicts of interest).

Theme IV addresses responsibilities to society as a whole and to more specific institutions within it. Its five principles encourage school psychologists to apply their knowledge and skills to promote healthy environments for children. School psychologists should be aware of federal, state, and local laws governing educational and psychological practices. They should contribute to the instruction, mentoring, and supervision of those joining the field or desiring to expand their skills, as well as to conducting and disseminating research. Finally, they should be self-aware of ethical conflicts they may experience and act accordingly, and should also provide peer monitoring to promote public trust in school psychology services.

Competence in Assessment

Theme II, of all the general ethical themes highlighted by the NASP, most directly addresses competence and responsibilities in assessment practices. Its first principle addresses competence. School psychologists are implored to reflect on their training experiences and to engage in only

those practices stemming from these experiences. They are also encouraged to participate in exercises to enhance their understanding and skills in working with students and families from diverse backgrounds, as well as to engage in professional development activities (such as attending professional conferences and self-study) to remain current in their knowledge and practices. In the same vein, the second principle encourages conscientious professional behavior and acceptance of the responsibilities that come with being a school psychologist. Specific reference is made to ensuring the accuracy of written reports, clearly communicating assessment results and other information, and monitoring the effects of recommendations and interventions. (See Chapter 8 for more information about preparing psychological reports and presenting information to parents and other caregivers.)

The third principle directly addresses assessment and intervention practices, and we highlight specific recommendations addressing assessment practices. One series of recommendations addresses measurement integrity and the fidelity of interpretations. For instance, school psychologists should rely on scientific evidence to select the optimal assessment instruments. In particular, they should select those with the strongest body of reliability and validity evidence supporting their use for the intended purposes (see Chapter 5). They should not violate the rules for uniform administration of standardized assessment instruments (see Chapter 4); if they do so, by error or by design through the use of test accommodations, they should report this in their oral and written descriptions of assessment results (see Chapter 8). Furthermore, they should use the most recently published and up-to-date normative data available (see Chapter 5) and exercise sound professional judgment in evaluating the results of computer-generated summaries of results and interpretive narratives.

School psychologists should rely on scientific evidence to select the optimal assessment instruments.

Theme II includes recommendations addressing general ideals of assessment practices. These ideals include broad and comprehensive assessments: (1) those that stem from multiple sources of information, including informants (e.g., teachers, parents, and students) and assessment instruments; and (2) those that represent all areas of suspected disability, such as health, vision, hearing, social–emotional functioning, motor abilities, and communicative status. When these assessments are complete, results should be presented in a clear and meaningful way (see Chapter 8). In addition, several recommendations for assessing children from culturally and linguistically diverse backgrounds are provided. For example, school psychologists are implored to conduct assessments that are fair for all those assessed. Thus, the process of selecting assessment instruments should include consideration of the potential examinee's disabilities and other limitations, as well as the examinee's cultural, linguistic, and experiential background (see Chapters 5 and 13). In addition, administration of assessment instruments and interpretation of their results should also take into account these characteristics of the examinee, to ensure that the results accurately represent the targeted constructs being measured. For example, school psychologists should promote quality control in the training and use of interpreters during testing practices (see Chapter 13).

The process of selecting assessment instruments should include consideration of the potential examinee's disabilities and other limitations, as well as the examinee's cultural, linguistic, and experiential background.

The next principle under Theme II addresses record keeping. In particular, school psychologists are implored to include only documented and relevant information from reliable sources in their records, which include psychological reports (see Chapter 8). Consideration of this standard should ensure that school psychologists report only information that is directly relevant to the referral concerns or eligibility decisions. For example, they should reflect on whether personal information about a child's immediate and extended family (e.g., histories of substance use or mental health problems) is relevant. In addition, they should ensure that they report the source of such information to support its truth value. The final principle under Theme II addresses intellectual property and copyright law, as well as record keeping, test security, and parents' and guardians' access to testing material (as previously discussed). It is particularly important to make sure that access to assessment results on computers and computer networks is restricted to authorized professionals. If this is not possible, then it is best to avoid the use of networked computers altogether and use the best personal computer security available. Psychologists should also assume that email is permanent and potentially public.

SUMMARY

Assessments should be grounded in the ethics of psychology. We find that with each additional reading of the APA's (2002, 2010a) Ethical Principles of Psychologists and Code of Conduct and the NASP's (2010a) Principles for Professional Ethics, we are inspired to do more than is commonplace and to be stronger advocates for children, families, and schools. We are also reminded of general and specific standards that we meet only partially or inconsistently.

On the one hand, we should aspire to ideals such as beneficence, justice, and integrity; strive for excellence in all aspects of our practices; and engage in lifelong learning to achieve and maintain expertise in professional ethics. On the other hand, we should be practical in applying day-to-day ethical practices involving assessment. We should remember that parents and children have the right to know the services we provide and the right to privacy; that assessment results are privileged and sensitive information; that our test content should remain secure to ensure its validity in assessment; and that we should strive to use the best assessment instruments and promote optimal testing environments to produce the most meaningful test results. We hope that the following chapters provide the information that will allow you, our readers, to reach these goals most effectively.

CHAPTER 4

The Assessment Process with Children and Adolescents

with Ryan L. Farmer

It is important to have a strong knowledge of the reasons for assessment, the typical steps in the assessment process, and the potential influences on test performance that can be controlled during standardized testing or acknowledged when interpreting test results. This chapter addresses these issues and provides practical tools that will promote the most accurate assessment of cognitive abilities through standardized testing.

THE COMPREHENSIVE ASSESSMENT PROCESS—AND HOW INTELLIGENCE TESTS FIT IN

Assessment is a broad term that refers to the process of collecting data to make informed decisions. Psychological and educational assessments are typically conducted for one of five reasons (and sometimes for multiple reasons). First, they are conducted for *screening* purposes—to rapidly identify those with specified characteristics, so that more in-depth assessment may be conducted. Screening children for hearing or vision deficiencies, for reading problems (via oral reading fluency probes), and for internalizing problems such as depressive disorders (via self-report rating scales) is relatively common. Intelligence tests (including brief or abbreviated intelligence tests; see Chapter 7) are often used to screen for intellectual disabil-

***Assessment* is a broad term that refers to the process of collecting data to make informed decisions.**

Ryan L. Farmer, MA, is a doctoral candidate at the University of Memphis. His research interests include intelligence assessment, threat assessment, and psychometric considerations in test selection.

ity (ID) and intellectual giftedness. Second, assessments may be conducted for *diagnosis* or *eligibility determination*. Assessment data may be used to determine whether a child or adolescent meets one of the criteria for a mental disorder as outlined in the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000, 2013), or such assessment may be prescribed by legislation (currently the Individuals with Disabilities Education Improvement Act of 2004 [IDEA, 2004]) to determine eligibility for special education services. Third, assessment may be conducted for *problem solving*—to identify problems, validate them, and develop interventions to address a concern. Fourth, assessments may be completed for *evaluation* purposes. For example, repeated assessments may be completed to determine the effectiveness of interventions (via progress monitoring) through the administration of fluency-based academic tasks. Finally, assessments may be completed as *indirect interventions*—because of their direct effects on the person being assessed—so that the process of completing the assessment leads to changes in the person and his or her behavior.

It is important to understand that *testing* is a more specific term than *assessment*, and that testing is typically only one component of an assessment. *Testing* refers to the process of collecting data via standardized procedures for obtaining samples of behavior, to draw conclusions about the constructs underlying these behaviors. A *construct* is an "attribute of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). In the case of intelligence tests, these constructs are cognitive abilities. When considering testing, most psychologists think of individualized standardized testing (i.e., tests administered by an examiner on a one-to-one basis with the examinee for screening, diagnostic, or eligibility purposes), and this book focuses on such tests. In contrast, most educators think about group-administered tests—those administered to a class or to all students in a school (as in end-of-the-year, high-stakes tests) for evaluation purposes.

Just as tests are only one component of the assessment process, there are many other components: reviews of records (e.g., report cards, prior reports and test scores, and incident reports); reviews of permanent products (e.g., completed class assignments); interviews conducted with parents, teachers, and other caregivers; systematic direct observations in classroom, home, or clinic settings; behavior rating scales completed by parents, teachers, and other caregivers; and interviews with and self-report rating scales completed by children and adolescents. Given the high degree of comorbidity (i.e., overlap) of academic and behavior problems displayed by children and adolescents, comprehensive assessments are often necessary to ensure that all relevant problems are evaluated (Frick, Barry, & Kamphaus, 2009). Best practice in assessment is to complete a multimethod, multisource, and multisetting assessment of children and adolescents' learning, as well as their behavioral and emotional functioning (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Whitcomb & Merrell, 2012), and such assessments are largely mandated for determining eligibility for special education. Thus the best assessments draw data from many different assessment techniques, from the child or adolescent being assessed as well as multiple knowledgeable persons who observe the child or adolescent, and from behaviors displayed across more than one setting (e.g., the testing session alone).

Intelligence tests are only one component—and sometimes an unessential component—of comprehensive assessments. Because intelligence tests produce measures of one of the most explanatory variables in all of the social sciences, psychometric *g*, we believe that they should be considered in seeking answers to children's academic problems. We do not claim, however, (1) that they should always be included in a comprehensive assessment or (2) that hours and hours of intel-

ligence testing and days of poring over its results are necessary in most cases. In fact, we recommend the following texts, which target other assessment instruments and methods that may be more useful in identifying the central problems underlying academic problems and developing effective interventions: Kamphaus and Campbell (2006); Chafouleas, Riley-Tillman, and Sugai (2007); Frick et al. (2009); Mash and Barkley (2007); Whitcomb and Merrell (2012); Sattler and Hoge (2006); and Steege and Watson (2009).

Because intelligence tests produce measures of one of the most explanatory variables in all of the social sciences, psychometric *g*, we believe that they should be considered in seeking answers to children's academic problems.

PRELIMINARY ASSESSMENT

Gathering Background Information

A thorough history of the child or adolescent being assessed should be obtained prior to administering intelligence tests. Background information is often needed to determine age of onset of the presenting problem (i.e., during which developmental period the problem surfaced); the course or prognosis of the problem (i.e., whether the presenting problem is becoming worse over time); etiology (e.g., whether the problem runs in the family), and the results of previous assessments or interventions. In particular, the following types of information should be obtained: referral information; demographic information (including age, grade, and race/ethnicity); family structure and history; the child or adolescent's medical history (including prenatal and perinatal history, genetic conditions, ear infections, head injuries, and surgeries and hospitalizations); developmental history (including speech–language milestones and motor milestones); social history (including major life stressors and traumatic events); and educational history (including grades, prior test scores, academic problems, behavior problems, and prior interventions). Although this information could be obtained by using a response form (a.k.a. paper-and-pencil format—e.g., the Behavior Assessment System for Children, Second Edition [BASC-2] Structured Developmental History Form; Reynolds & Kamphaus, 2004), this information should be obtained in an interview if possible. Interviews provide the opportunity to follow up on issues raised by informants, and they also facilitate rapport building. (The working relationship between an assessment professional and a client during assessment is often called *rapport* in the professional training literature.) Your rapport with informants will promote their engagement in the assessment process and facilitate your clear communication of results after the assessment is complete.

Screening

In addition, a thorough screening for possible impediments to testing should be conducted. Throughout this chapter, we consistently address methods to identify and prevent or minimize *construct-irrelevant* influences on test scores (AERA et al., 1999). These confounds as they apply to validity evidence are described in more detail in Chapter 5, but in this chapter we address them as relevant to individual testing cases. According to Bracken (2000),

> An assumption made about the psychoeducational assessment process is that examiners have made every effort to eliminate all identifiable construct-irrelevant influences on the child or ado-

> lescent's performance and the resultant test scores. That is, the goal in assessment is to limit assessment to only construct-relevant attributes (e.g., intelligence), while limiting the influence of construct-irrelevant sources of variation (e.g., fatigue, lack of cooperation, emotional lability). Before important decisions can be made with confidence about a child or adolescent's future educational plans, possible treatments, or medications, examiners must be comfortable with the validity of assessment results. Only when all construct-irrelevant sources of variation have been eliminated or optimally controlled can examiners attest to the validity of the assessment results. (p. 33)

Basic methods can be used to screen for problems with sensory acuity, speech and language, motor control, and behavior as potential confounds. Based on our review of the literature and consultation with professionals in such fields as optometry, audiology, speech–language pathology, and physical therapy, we have identified items that parents, teachers, or both should complete as part of the initial stage of assessment (American Speech–Language–Hearing Association, n.d.; Centers for Disease Control and Prevention, n.d.; Gordon-Brannan, 1994; Colour Blind Awareness, n.d.; Teller, McDonald, Preston, Sebris, & Dobson, 2008; Mathers, Keyes, & Wright, 2010; Suttle, 2001). For this book, we have created several assessment tools that are collectively called the Screening Tool for Assessment (STA).

Form 4.1 is the STA Parent and Caregiver Screening Form, and Form 4.2 is the STA Teacher Screening Form.* These forms include items addressing visual acuity problems (items 1–6), color blindness (items 7–10), auditory acuity problems (items 11–16), speech and articulation problems (items 17–22), fine motor problems (items 23–29), and noncompliance (items 30–32). If the patterns of item-level responses indicate serious concerns (e.g., if more than half of the items are marked "Yes"), you should delay testing and inquire about the extent of these apparent problems, discuss whether prior screenings have indicated problems, or consider whether more thorough screening should be conducted. If problems in these areas cannot be corrected before an assessment (e.g., vision problems corrected with glasses), you should (1) select tests so that they are not influenced by them or (2) develop test accommodations to implement during testing. Overall, it is important to consider these influences and eliminate them, because they are potential confounds that will undermine the accuracy and meaningfulness of your test results.

THE TESTING PROCESS

In this section, we address how to prepare yourself and the testing environment for testing, how to establish rules and expectations for the testing session, how to build rapport and interact with children and adolescents during test administration, how to observe test session behaviors, and how to judge the "validity" of the test results.

Preparing for Testing

As you prepare for the testing session, you should be physically and mentally ready to test; appropriately arrange the testing materials and adjust the testing environment to maximize efficiency

* All forms appear at the ends of the respective chapters.

and eliminate potential confounds; and have a thorough grasp of the test's administration procedures (e.g., rules for subtest stopping points and querying).

Dress, Accessories, and Equipment

Because most testing sessions last approximately 2 hours, you should dress comfortably yet professionally. Although most sessions will be completed at a table with you sitting in a chair, you may need to sit on and crawl about on the floor when you are testing young children. Dressing in comfortable slacks and loose-fitting clothes is important (see Bracken, 2000, for more information about testing young children). In addition, you should be mindful of how your accessories may interfere with testing. For example, decorative rings and bracelets should be avoided, and discretion should be used in selecting ties, necklaces, and earrings. These personal items should not become distractions during testing.

In the same vein, we discourage use of cell phones during testing. Although we know psychologists who use their cell phones for timing during testing, the risk is too great for cell phones (especially smart phones) to cause distractions by vibrating, ringing, beeping, or otherwise providing audible notifications. In addition, it is common for children to use their parents' cell phones for games and other activities, so they may see your phone as a game console rather than a part of the testing equipment. We also occasionally have seen psychologists using digital wrist watches for timing. Unless such a watch is removed from the wrist, using it will be cumbersome; furthermore, wrist watches are often somewhat difficult to manipulate because of their small buttons. We recommend using "old-school" athletic digital stopwatches or digital kitchen timers. These devices should fit in the palm of the hand, should count up (and not only down, as some timers do), and should have beepers removed so that they operate silently.

It is becoming increasingly common for both intelligence and achievement tests to require tape players or CD players for administration. One example is the Woodcock–Johnson III (WJ III) battery (Woodcock, McGrew, & Mather, 2001). In general, you should (1) select the highest-quality equipment (e.g., a tape player with good speakers and a counter), and (2) test this equipment before initiating testing. Headphones (a.k.a. ear buds) are also a necessity, especially with tapes that require items to be queued up. Using CDs and CD players is far easier than using audiotapes and tape players, because you can move across items without using headphones or a counter.

Finally, you should have other materials at your disposal. Of course, numerous pencils are needed, and you will need to consider variations in the types of pencils required across tests. For example, erasers are prohibited during administration of some tests; they should be removed before testing. In addition, #2 lead pencils are typically called for, but sometimes red pencils are required. You should also maintain a collection of tangible incentives. For example, small stickers may be useful to reinforce attentive and compliant behaviors during testing. For young children, small food items (e.g., M&Ms and raisins) may also be helpful for the same purpose. Issues pertaining to the use of such incentives are addressed later in this chapter.

Selecting and Arranging the Testing Environment

You should select a testing environment that has minimal distractions and is otherwise ideal for testing (Bracken, 2000). The room should be well lighted, be kept at a comfortable temperature,

and have adequate ventilation. It should contain furniture that is appropriately sized for the child or adolescent being tested. Chairs should be short enough so that a child's feet can touch the floor, and if they cannot, consider placing a box or large books under the child's feet (Kamphaus, 2001). Tabletops should have smooth surfaces, and children should be able to reach them without straining.

You should also consider where you and the child or adolescent being tested should sit. Although the pattern of seating (e.g., sitting across from examinees or beside them) is dictated by the test's standardized procedures, you should strive to seat yourself at the long end of a rectangular table (if possible), in order to arrange your materials more easily and to allow for more writing space. Circular tables are not ideal for testing when (1) they are so wide that you cannot comfortably reach across the table, or (2) it is difficult to space testing material appropriately from the edge and to judge the correctness of some responses when there is no straight edge. Also, seat the child or adolescent so that he or she is not facing distractions (e.g., a window with a view of the playground), and consider arranging the seating so that you are between the child or adolescent and the testing room door.

You should consider the time of day for testing. In elementary school settings, it may be that testing in the morning is ideal; in high school settings, the opposite may be true for adolescents, who may be groggy in the morning. In clinic settings, weigh the costs and benefits of testing at the end of the school day or in the early evening versus having the child or adolescent miss school to complete testing.

Finally, you should reserve enough time to complete each test in its entirety for two reasons. First, most tests were normed that way (see Chapter 5). Second, completing a test in one session prevents you from having to use different norms to score subtests if the delay is extremely lengthy. (In general, if the delay between sessions is no more than 2 weeks, use the child or adolescent's age at the time of the first session unless other recommendations appear in the test's scoring guidelines.)

Knowing the Test and Preparing Testing Materials

We cannot stress enough the importance of your being prepared for testing. It is vital not only to prepare yourself for testing, bring the right support materials, and select and arrange the testing environment, but also to know the administration and scoring procedures well and to prepare the testing materials for an efficient administration. Sattler (2008) has conveyed the importance of preparation for testing—especially for psychologists in training:

> Your goal is to know your tasks well enough that the test administration flows nicely, leaving you time to observe and record the child or adolescent's behavior. To do this you will need to have learned how to apply the administration and scoring rules, how to find test materials quickly, and how to introduce and remove the test materials without breaking the interaction and flow between you and the child or adolescent. Do not study a test manual as you give the test, because doing so will prolong the testing, may increase the child or adolescent's anxiety, and may lead to mistakes in administration. However, in most cases you must refer to the directions and scoring guidelines in the test manual as you administer the test. Most individually administered tests do not require you to memorize test directions. Use highlighters, adhesive flags, index tabs, and other aids to facilitate swift and efficient use of the test manual. (p. 201)

When you are first learning to administer a test, it can be daunting to remember all the rules of standardized administration while maintaining rapport, administering and scoring items, and observing test behavior. To help you reach these goals, you should mark up your test records in advance of testing. Start points can be circled; varied rules for discontinuing can be highlighted; and unique queries, pronunciations of items, and reminders to prevent common errors can be written on sticky notes or in blank spaces on test records. These markings will lessen the amount of information you need to juggle in your mind during testing.

In addition to knowing the intricacies of the tests, it is also useful to prepare your test materials (opening manuals to the first page, setting up easels, setting out pencils, organizing manipulatives for presentation, cueing up tapes for auditory administration, etc.) before initiating the testing session. Of course, if you are pressed for time, these activities can be completed while building rapport; however, this dual tasking is not ideal, because it does not allow you to exert optimal attending skills during the early stage of rapport building.

What should you do if parents request that they be able to observe your testing session by sitting in the testing room with you? It is not a good idea to allow parents to do so if the child or adolescent is age 4 or older (see Bracken, 2000). We have found that most parents understand and respond appropriately when it is explained that the tests were not developed to be completed with a parent in the room, and that their being in the room may be disruptive to the child or adolescent and lead to lower scores. Kamphaus (2001) has suggested that parents should also be told that they should consider how stressful it may be (for them!) to observe the assessment: "Parents want their child to do well, and when they do not, it can be extremely punishing to a parent, especially when they know that their child is taking an intelligence test" (p. 98). If a parent insists on being in the room, or if the child will not separate from the parent despite coaxing, the parent can sit in the testing room outside the child's line of vision (i.e., behind the child). The parent should be instructed (1) to stay silent during testing and (2) not to reveal the content of test items after the session is complete (in order to maintain test security).

Beginning Testing

First Contact and Initial Rapport Building

Because most intelligence tests require one-on-one administration, you will need to retrieve the child or adolescent from a classroom or other school setting (for a school-based assessment) or from a waiting room (for a clinic-based assessment) to initiate testing. It is important to begin developing rapport from the initial contact and to maintain rapport throughout the testing sessions. This rapport is likely to be relatively short-term—across fewer than four sessions and perhaps an additional meeting to describe the results. Your goal in fostering rapport is to obtain information about the child or adolescent, versus to facilitate verbal discussion, reflection, and initiation of therapeutic techniques and strategies for addressing psychological problems (as in counseling and psychotherapy). Thus, the goal of rapport building and rapport maintenance during testing is to ensure enough of a bond with the child or adolescent so that he or she (1) does not feel negative emotions (e.g., fear or worry) that might undermine the assessment, and (2) is motivated to respond to test items and questions in a manner that represents the targeted areas well. In doing so, you want to consider the child or adolescent's age and find the appropriate balance between being approachable (fun, interesting, and humorous) and being more formal and businesslike (Bracken, 2000).

Before you meet the child or adolescent for the first time, consider information you may already possess that may aid you in generating topics of conversation. For example, you may have already gathered information from parents (via their completion of initial interviews or developmental history forms) and from direct observations of the child or adolescent. In addition, you may have scanned materials posted outside the classroom and learned something about the child or adolescent. From these sources, you may have learned about topics or people on which recent classroom projects have focused. In addition, for school-based assessments, it is wise to consult briefly with the child or adolescent's teacher to ensure that you are retrieving the child or adolescent during a segment of the day that does not cause major disruption of important activities in the classroom or during activities that the child or adolescent finds most enjoyable, such as those during support classes. If the timing is not right, you can return to your testing room, engage in other activities (such as scoring protocols or emailing colleagues), and then return later to retrieve the child or adolescent.

When you first meet the child or adolescent, smile, greet the child or adolescent in a friendly manner, and use his or her name. We suggest offering your hand to shake, and, with young children, squatting to be roughly at their eye level when you first meet them. Also, keep the greeting brief; you will have more time to chat and explain your goals soon afterward. You might say, "Hi, Erica. My name is Ms. Barker. I am here to work with you today and talk about school. Please walk with me to my room." We admit to falling prey to a strategy that, on occasion, backfires. Consistent with guidelines for securing a child or adolescent's assent during research projects, we often ask, "Would you like to come with me?" at the end of our greetings. The occasional problem is that children or adolescents may respond negatively, saying that they would rather not do so. When such a response occurs, it lengthens the greeting process, but it may also allow the child or adolescent to express reasons for hesitations that you can address.

As you begin to converse with the child or adolescent while you walk toward the testing room, continue your rapport building by asking open-ended questions about "easy" topics (hobbies, how the day is going, any fun activities the child or adolescent has completed recently, etc.). We find discussion of favorite colors or favorite letters (for young children), pets, and sports to be particularly easy for many children. As you facilitate conversation, strive to be engaging and enthusiastic about the child or adolescent's responses and experiences. Use extenders (e.g., "Oh" and "Umm") and brief positive feedback (e.g., "That's cool" and "I like that, too") in response to the child or adolescent's statements, so that you do not monopolize the conversation. Most children and adolescents appreciate this focused attention from an adult.

Addressing the Purpose of Assessment and Confidentiality

Once the child or adolescent is in the testing room, and you have silenced your cell phone and placed a "Do Not Disturb" sign on the door, you can provide more details about your goals for the testing session. Although excellent scripts are offered in Sattler (2008) and Kamphaus (2001), use a script like this one that is appropriately modified to address important issues related to the assessment. You might begin this way:

> "I met your mother/father and your teacher, and now I want to get to know you. They tell me that you have been having a hard time at school. I want to know more about what you have already learned at school and how you go about learning new things, and I want to talk about

what we can do to make school better for you. Our work today should last about 2 hours. Is this a good plan to you?"

Most children and adolescents will nod their heads. Then you might ask the child or adolescent to repeat to you what you have said. Say, "So tell me what we're going to do today," and accept any reasonable summary.

Consistent with the legal and ethical guidelines discussed in Chapter 3, it is important to explain confidentiality and its limits. Although we do not believe that these issues are commonly addressed with children or adolescents before testing sessions, it is best practice to do so. You might say,

"We need to follow certain rules when we work together today. I want you to feel that you can talk to me, so I'll keep what you tell me between you and me [or private]. That means that I won't go telling your friends, anyone outside of school, or anyone else besides your parents and teacher[s] about our work together. What I share with your parents and teacher[s] will be things to make school better for you. How does that sound?"

Then you can explain the limits of confidentiality and its limits:

"Again, I will keep our work together private unless several things happen: (1) You tell me that someone (like an adult) has hurt you or someone you know; (2) you tell me that you are going to hurt someone; (3) you tell me that you are going to hurt yourself; or (4) if a police officer or court judge requires me to share the information. When these things happen, you and I would need to talk more about them and share them with your parents, teacher[s], and other adults who care about you. What do you think about this part?"

You can close the introduction to the testing session by summarizing, adding other rules, encouraging communication with you, and perhaps inserting humor as a bridge to introducing the test. You might say,

"Other than these things, what we talk about stays between you and me, because I want to help make school better for you. It is very important that you try your best and that you are honest with me because we need to figure things out together."

You may want to add other rules, depending on the child or adolescent's age. For example, for young children you may add rules about staying in the room and asking your permission before touching objects in the room. For others, you may also encourage them to communicate their discomfort at any point during the testing session. Finally, you can consider lightening the mood by telling a joke or by saying something obviously incorrect. For example, some of our favorite strategies in this vein are calling a child by the wrong name (e.g., "OK, are you ready to move on, Esmeralda-ina?" and "So I've forgotten. Your name is Eli . . . Manning, right?") and indicating that the child is much older than his or her actual age (e.g., "I have it written here that you are 10 years old. Is that right?" and "So is it correct that you are in the ninth grade?"). Almost all children will correct your errors. If these questions are asked with a wry smile, we have found that children tend to find them a bit silly; when you say, "Oh, that's right!" and shake your head, throw up your arms,

and look exasperated by your error, this technique aids in rapport building. You should develop your own strategies like these to lighten the mood.

Introduction to Testing and Screening for Confounds

Before you begin testing, it is important to describe your expectations for testing and to screen for potential sensory deficits and other personal influences on test performance, in order to rule out potential construct-irrelevant influences on performance on the day of the testing.

Most intelligence tests have incorporated introductions to testing in their standardized procedures, but we encourage you to review them carefully and enhance them with some additional content to promote clear expectations. Essentially, you want to ensure that the children and adolescents understand at least four different components of the testing: (1) that they will complete various tasks; (2) that some of these tasks' items will be easy and others difficult; (3) that they should exert effort and persistence to do their best; and (4) that is it acceptable to report that they do not know answers. We tend to read introductory directions for most intelligence tests and supplement these with "It's OK to ask questions, guess, or say, 'I don't know.'"

Before you begin testing, it is important to describe your expectations for testing and to screen for potential sensory deficits and other personal influences on test performance.

Ice Breakers and Screening

From our experience and review of the literature, we understand that it has historically been very common for psychologists to administer simple tasks as "ice breakers" before beginning testing. Drawing tasks, such as the Bender–Gestalt (Bender, 1938) and the House–Tree–Person (Buck & Warren, 1992), were often administered prior to intelligence tests in years gone by. One goal of their administration appears to have been engaging a child or adolescent in some low-stress, paper-and-pencil task before beginning to administer the intelligence test items, which are often orally administered. These tasks provided another buffer between entering the testing room and beginning testing. In some cases, these tasks also allowed examiners to screen directly for potential problems that may undermine the validity of the test scores. However, in our review of the literature, we found no ice-breaking activity that accomplished all of these goals—especially effectively screening for the variety of problems that may undermine accurate testing. To accomplish these goals, we developed the STA Direct Screening Form to accomplish this goal (see Forms 4.3 and 4.4).

The STA Direct Screening Form begins with general questions targeting emotional states and preparedness for testing. The first section covers global impressions of health and mental state, the prior night's sleep, feelings of confidence, and motivation. The next section includes sections devoted to screening for sensory deficits and fine motor control problems. These questions are not intended to be administered in a rigid fashion; instead, they should guide your brief screening interview. Be sure to follow up responses with appropriate extenders, paraphrasing, more specific questions ("What do you mean?"), and "soft" commands (e.g., "Tell me more about it"). There is no reason to complete additional screening if all evidence converges on the absence of sensory deficits and fine motor control problems, but direct assessment of these issues may be useful. The

remainder of the STA Direct Screening Form contains assessment items targeting sensory deficits, fine motor control problems, as well as articulation problems and intelligibility.

First, the STA Direct Screening Form includes items targeting visual acuity. Children and adolescents are asked to name letters printed on a page to screen for potential vision problems that may interfere with the test administration (see Form 4.3 for instructions and Form 4.4 for items). The letters printed on the top line should be large and distinct enough for every examinee to see, and corrective feedback can be used to train young or low-functioning examinees. It is essentially a practice trial. After the completion of this trial, you should ask examinees to read the rows below it. Because no intelligence test uses type smaller than the second row of letters (from the top) in Form 4.4 (i.e., 12-point type), vision screening is passed if the examinee accurately reads seven of the nine items in any row below the top row. Do not administer the vision screening items to those (especially young children) who have not yet mastered the names of all letters of the alphabet.

Second, direct assessment items target color blindness. Examinees are asked to identify the colors of six squares (see Form 4.3 for instructions and Form 4.4 for items). They should be able to identify all six colors, and errors in identifying the red and green squares will be especially diagnostic for color blindness in those old enough and high-functioning enough to know color names. Do not administer the color blindness items to those who have not yet mastered the names of the basic colors. Third, direct assessment items target fine motor skills (see Form 4.3 for instructions and Form 4.4 for items). Examinees are asked to trace a horizontal line, a star, and a circle, and to write a simple sentence. Do not administer the item requiring writing to those (especially young children) who have not yet mastered letter printing. Consider deviations from the lines and malformed letters, as well as unsteady pencil grips, hand tremors and jerky movements, impulsive and messy responding, and signs of frustration when an examinee is completing these items, as these may be associated with fine motor control problems. Responses are judged qualitatively; the screening is passed if the examinee responds with reasonable accuracy in tracing or the writing is legible to a stranger who would not know what the sentence in the last item should say.

The remaining direct assessment items from the STA Direct Screening Form do not require visual stimuli (as those included in Form 4.4 do). The next items target auditory acuity (see Form 4.3). These screening items require the examinee to play a listening game, where you give brief commands and the examinee points to parts of his or her face and body (Howard, 1992). During Part A, administer all the items in a slightly louder voice than typical to establish a baseline understanding of the commands, but during Part B, ask the examinee to close his or her eyes while you readminister the items in a whisper. During Part B, observe behavioral signs of hearing problems, such as turning the head to favor one ear and delayed or vague responding (e.g., hovering of the hand near their head). The screening is passed if the examinee responds accurately to at least five of the seven items in Part B.

Finally, direct assessment items target articulation and intelligibility (see Form 4.3). The child or adolescent is asked to close his or her eyes and to repeat words you have stated at a normal volume. These items were developed for the STA on the basis of Shriberg's (1993) findings regarding the development of more challenging phonemes in children. Phonemes included in each word represent the so-called "Late 8" sounds (i.e., the last eight sounds that children are expected to acquire in speech). Phonemes are represented in three positions (initial, medial, and final) across the words (when possible), because these phonemes present different levels of difficulty at each position. For our purposes, we have used the International Phonetic Alphabet when sounds match

the English alphabet letters (e.g., /l/). However, when appropriate phonetic symbols would not be commonly recognized (e.g. ʃ representing the *sh* sound), a representation of that phoneme appears in quotation marks. When you are considering patterns of responses, keep in mind that 75% of children should acquire the phonemes /l/, /z/, /s/, *sh*, and *zh* by age 5. The phonemes *th* (voiced, as in the word *although*, and voiceless, as in the word *think*) and /r/ should be acquired by age 6 (Shriberg, 1993). On the form, circle incorrect phonemes; the screening is passed if the examinee provides at least 20 of the 23 correct phonemes.

In addition, a brief rating of intelligibility is available that coincides with items on the STA Parent/Caregiver and Teacher Screening Forms (Forms 4.1 and 4.2). *Intelligibility* is best defined by Bowen (1998) as the percentage of an individual's spoken language that can be readily understood by an unfamiliar listener. Flipsen's (2006) intelligibility formula is as follows:

$$\frac{\text{Age in years}}{4} \times 100 = \%\ \text{understood by an unfamiliar listener}$$

As such, a child's intelligibility should increase by 25% for each year after birth. A 1-year-old is expected to be 25% intelligible to strangers, whereas a 3-year-old is expected to be 75% intelligible to strangers. Taking comparable ratings from familiar listeners (e.g., parents) and unfamiliar listeners (e.g., you as the examiner) results in a clinical, albeit subjective, accounting of intelligibility in children. For the purposes of assessment, low intelligibility can affect a child's ability to respond verbally to items. As such, low intelligibility ratings should be considered potentially invalidating when your examiner rating is below 75%. In these cases, it may be wise to consider using language-reduced composite scores or multidimensional nonverbal intelligence tests.

Interacting with Children and Adolescents during Testing

Once you are reasonably confident that the examinee has no sensory acuity, fine motor, speech, or other personal problems that will undermine testing, introduce the first test items. In general, begin testing when the child or adolescent appears ready—and sooner rather than later.

In general, begin testing when the child or adolescent appears ready—and sooner rather than later.

Standardized Procedures

Following standardized administration procedures, once these are studied and practiced repeatedly, should allow for a brisk administration. In addition, we cannot stress enough the importance of following standardized procedures with almost rigid adherence. Kamphaus's (2001) point about this is spot on: "If an examiner does not use standardized procedures, then he or she should not use the norms tables" (p. 106). Although it is possible to interpret intelligence tests from a qualitative and idiographic perspective (see Chapter 7), the best-validated findings come from interpretation of norm-referenced scores. We agree with Kamphaus's point that psychologists in training should think about their own experience of taking group-based standardized tests, such as the ACT, the SAT, and the GRE. You should consider all the rules and decorum (the proctors, the reading of scripts, and the strict time limits) that are central to the standardized testing enterprise, and strive

to replicate such strict adherence. You should read instructions, item introductions, and feedback after errors exactly as they are written—even if you are motivated to improve the examinee's understanding. For example, Sattler (2008) has implored examiners to "never ad lib, add extraneous words, leave out words from instructions or the test questions, or change any test directions because you think that the altered wording would improve the child's performance (unless the test manual permits changes)" (p. 203).

Accurate timing of items on intelligence tests is vital to measuring individual differences in cognitive abilities, and we have five points to offer about timing. First, there is no reason to hide your stopwatch or otherwise be surreptitious in timing. We agree with Kamphaus (2001) that you should be natural about timing and use of the stopwatch. Now, perhaps more than ever, children and adolescents are accustomed to completing timed (and fluency-based) tests; it is unlikely that they will be unnerved by your using it. Second, timing should begin once you have completed the instructions. Once timing for an item or subtest has been initiated, you should not stop timing until the item is complete or the time limit has ended. Do not stop timing when children or adolescents ask questions, when they make errors in response to individual items, or when you must prompt to maintain standardization. Third, discontinue the items or subtests when the time limit has expired, or record the exact completion time (typically rounding down to the nearest second vs. recording minutes, seconds, and milliseconds). Fourth, score the last response provided within the time limit. Most subtests have you award no credit for items completed after the time limit, although there are a few exceptions in which partial credit is given for incomplete performance. Finally, do not share the time limit or the time remaining with the child or adolescent, even if asked. If you are asked a question during a timed trial, you might make eye contact and calmly say, "Just keep working," and then break eye contact so as to not engage in conversation.

We are seeing a trend toward more explicit description of "soft time limits" targeting test items—especially items requiring expression of knowledge and oral expression. In general, you should present the next test item after allowing a child or adolescent an appropriate amount of time to respond to a difficult question. However, some tests require that a response to an item be given in about 20 or 30 seconds, to enhance the efficiency of testing. Other tests, as noted previously, apply an absolute time limit to item-level responses. You should find the right balance between (1) allowing examinees to continue working on responses after the time limit has expired, and (2) stopping them in development of a response and ushering them to the next item. It is important, however, to score the response that was available before the time limit expired.

Extending and Clarifying Responses

During testing, you may need to encourage the examinee to extend or clarify an initial response. Typically, you should offer *queries*, such as "What do you mean?" and "Tell me more about it." Although these queries are very similar across intelligence tests, you should always use the queries prescribed by each test. We recommend using sticky notes or the like in the test record to remind you of these queries when you are administering several different tests as part of an assessment. For most intelligence tests, the administration manual provides sample responses that must be queried, but you should also query vague responses, regionalisms, or slang responses that are not clearly incorrect. However, you should be careful not to query any clear-cut responses that contain enough information to score without a query. Furthermore, you should record on your protocol the points when you queried (using Q to represent a query).

In general, you should consider all parts of a response—those parts that are offered before the query is issued, and those offered after the query is issued. In other words, score the whole of the response. For example, when asked about the meaning of the word *colander*, the child offers "You use it in the kitchen with water," and after you query the response, he or she offers, "It keeps you from pouring hot pasta down the drain," you would consider both parts of the response, and the child would earn as many points as if he or she had combined both parts into one sentence. However, it gets tricky when the response after the query is mediocre or outright flawed. If you issue a query, and the child offers a weak response or otherwise fails to improve the initial response (e.g., "Colanders are plastic and white"), score the quality of the initial response. However, if the query reveals a fundamental misconception—a totally flawed understanding—about the content of the item (e.g., "Colanders chop up food into pieces"), the initial response, regardless of its quality, is spoiled, and the item is scored 0.

Dealing with Audio Players

Using an audiotape player or CD player introduces challenges into the testing session that were not commonly experienced by those administering intelligence tests in the past. In addition to needing one more piece of equipment, if you are using an audiotape player, it is important to use a player that has a counter so that you can link your subtest start points to items on the tape. Doing so requires preparation, a set of headphones (if adjustments must be made during testing), and a good deal of rewinding and fast-forwarding. A CD player largely eliminates these problems, because most start points can easily be located.

When you are using any type of audio player, it is important to look away from the examinee (e.g., at your test record) when the item is being played, and then look expectantly at the examinee once the item has been completed. In addition, it is important to stop the CD or pause the tape (when allowed) to allow examinees who are slow to respond more time to formulate a response. Although there should be enough time between items for most children and adolescents to respond, we recommend (1) being conscious that the next item may be presented before a response can been offered, and (2) stopping or pausing before such incidents occur.

Scoring and Recording Responses

You should strive for a brisk administration, while also taking your time to score responses concurrently with the administration. If you cannot decide on a score for an item, leave it blank, mark it with a question mark, continue to complete the last item, or meet the discontinue rule while not including this item; score the item in question at a later time. When you are scoring, both example responses and general scoring criteria should be consulted, but keep in mind that the examples listed in the test manuals are guides to scoring and that some professional judgment (a.k.a. "real thinking" on your part) may be required. If you are a beginning tester or are learning a new test, we strongly encourage you to have a second party review your item scoring and raw score calculation. We are also strong advocates of using scoring software to prevent errors when the norm tables presented in manuals are consulted.

It is important to score all items according to the examples, general criteria, and scoring guides provided in test manuals, and to avoid applying subjective judgments that "bend the rules" for examinees. However, as just noted, professional judgment is needed—especially in cases

where the examinee responds in a manner that is more sophisticated than the sample responses. For example, an adolescent may respond to a picture vocabulary item by reporting the genus and species of a beetle, rather than referring to it as an "insect" in general or a "beetle" more specifically. Some children may calculate the exact number of weeks in a year (yielding a decimal fraction), versus telling you that there are 52 full weeks in the year. In such cases, it is incumbent on you to search books or the Internet, or to do the calculations yourself, to determine whether the response is correct or incorrect. In the same vein, you should not penalize the child or adolescent for mispronunciations due to articulation problems or to regional or dialect differences.

It is difficult to know how to score initial responses that include both correct and incorrect components. Although you should always consult the test manuals for specific guidelines, here are two commonalities across them. First, if several responses to an item are given, ask which one the child or adolescent wants you to score. You can say, "You said *X*, and you said *Y*. Which one is it?" or the like. (Note that some intelligence test manuals encourage you to score the last response given.) Second, for many tests, if the child or adolescent provides numerous correct responses of varying quality and the item is scored on a scale, score the best response.

When administering the tests, you should record verbatim responses to the degree possible. Record them in the lines on the test record or beside the item. However, exercise some discretion in selecting only the core of the responses to record. For instance, you need not record the initial stem of a response (e.g., "I think that an encyclopedia is a . . . "), interjections (e.g., "Darn!"), or idiosyncratic responses (e.g., "I think we learned that in third grade" or "That is a tough one") that are not clinically meaningful. Thus, extraneous content need not be recorded. It is also important to be able to use abbreviations when appropriate, but never place them in the item-scoring column.

Breaks, Check-Ins, and Encouragement

It is important to monitor the child or adolescent's motivation and persistence during testing, and to encourage full responses to all items. Throughout the testing session, you should "read" the child or adolescent's behavior, including facial expressions, posture, and verbalizations, to determine his or her level of motivation and fatigue (Kamphaus, 2001). Although you typically want to keep small talk to a minimum during subtests, we recommend "checking in" between subtests with questions like "How are you feeling?" and "Ready for the next one?" to gauge motivation and fatigue. We encourage energetic, brisk administrations of the tests, but for some children and adolescents, you may call for breaks every 30 minutes or so. You must consider, though, that (1) some children and adolescents have a difficult time readjusting to the testing session after a break; and (2) in a school setting, breaks lengthen the assessment session and potentially prevent exposure to valuable instructional content in the classroom setting.

Engaging in strategies to encourage full expression of knowledge and skills during testing is vital. Although you are able to provide encouraging feedback after correct responses to sample items—and some tests do, in fact, allow for such feedback after responses to test items—you should generally strive to give no indication of whether responses to test items are correct or incorrect. In particular, you should not indicate the quality of the examinee's responses through your facial expression or your verbalizations. We encourage our students in training to practice their "stone face with a smile" expressions throughout testing, and to develop nonevaluative verbalizations that can be consistently used. For example, we believe that saying "OK" or "That's fine," and nodding after responses, are both acceptable behaviors—as long as examiners are diligent about

using them for both correct and incorrect responses. Others may assert that presenting the next item signals clearly that the last response has been accepted, but we find that, with preschool- and elementary-school-age children, some other way of accepting their response seems warmer and more responsive.

We recommend two types of feedback designed to praise a child or adolescent's effort during testing. First, use various comments to recognize effort. Examples include "Thanks for your hard work," "I like how you thought through that one," "Excellent effort! Keep it up!" and "Keep giving it your all." We also encourage well-timed general comments, such as "You are doing a great job," and "Good work," that cannot be linked to correct responses and that are intermingled with comments about effort. Second, we encourage physical gestures and brief physical contact, such as fist bumps. We admit to engaging in some seemingly hokey behaviors, such as prompting children to give us "high fives" and offering "thumbs-up" gestures paired with phrases like "Good job" and a smile. Of course, these behaviors must be moderated in keeping with the child or adolescent's age. For example, preschoolers tend to enjoy silliness from adults, but the same cannot be said for many adolescents. In particular, adolescents may be particularly sensitive to insinuation of false praise or to repetitions of platitudes.

The use of tangible and edible incentives during testing is a topic of frequent debate (see Fish, 1988). We agree with Sattler (2008) that such incentives should not be used under normal circumstances; they may introduce unnecessary complexity and potential hazards into the testing situation (e.g., activating food allergies and choking, in the case of food items). However, we have seen benefits of the use of incentives with preschool-age children, children with ID, and children with attention-deficit/hyperactivity disorder (ADHD) when they are used to reinforce compliance and responding to items. We offer these recommendations. First, if you plan to use food items, obtain the parent or caregiver's permission to do so. Many food items, such as M&M candies or marshmallows, may be prohibited by the family. Obtaining such permission in school settings is often impractical. Second, food items often interfere with testing because they have the potential to make a mess (e.g., tortilla chip remnants on the testing table or chocolaty saliva running down a child's mouth). Food items are probably best reserved for preschoolers, but for older children and adolescents, we recommend that a token economy or point system be implemented so that they earn rewards for compliance (e.g., sitting in the chair and attempting as many questions as possible) that can be traded in at the end of the testing session for valued items (e.g., trading cards, stickers, and colorful pencils).

Monitoring the motivation and persistence of children and adolescents, and encouraging them to respond fully requires guiding them to work though the most difficult test items. Although you may have already shared with them that they can report that they do not know some answers (and that no one is expected to be able to answer all items), you should encourage them to guess by reminding them with "It's OK to guess" or "It's OK to answer even if you are not sure," and encouraging them to "Give it a try" or "Give it a shot." After using these techniques (and driving home the point that guessing is not a problem) and reading their behaviors as the testing progresses, you may also ask, "Want to give it a try or move to the next one?" or "Want to pass on this one?" Furthermore, you should be sensitive to their perceptions of failures, and remind them that effort is more important to you than correct responses. There seems to be some benefit of reminding children and adolescents that some items were developed for older adults. You may say, "Remember I told you that some of these items would be hard; they must be for moms and dads, huh?" or "That was a tough one, wasn't it? I like how well you worked on it, though."

Responding to Questions

Some anxious or inquisitive children and adolescents may barrage you with questions. Others, due to some of their limitations, may require you to respond to pleas for assistance. For example, some may ask you for feedback on items, asking "Was I right?" or "How am I doing?"; others may ask you for the answers to questions that they could not produce. Your standard responses to these issues should be to say, kindly, that you cannot tell them how well they are doing or what the correct answers are. Comments such as "You're doing fine, and I appreciate your hard work," and "Your answer sounds good to me," work well. We often refer to our testing rules to justify our inability to share the correct answers. We say, "I'm sorry that I can't tell you; it's one of the rules I must follow in doing my job."

Unless the test item targets an ability that would be enhanced by repeating the item (e.g., subtests targeting Short-Term Memory or Listening Ability), you can repeat questions or items upon request, but when you do so, it is important that you repeat it in its entirety so that you do not parse its components for them. For example, on a subtest targeting quantitative reasoning that requires children to calculate the difference in miles between two cities, a child might ask, "Did you say in miles?" Rather than replying in the affirmative, you should repeat the whole item to require the child to parse the important information from it. You may also repeat items when a response suggests that the child misheard or misunderstood words used in the item.

Observing Behaviors during the Test Session

As we have noted earlier in this chapter, you should observe test session behaviors to judge how these behaviors might undermine the validity of your score results. Even if you have gone through all the efforts to eliminate or to control optimally for construct-irrelevant sources of variation in test scores, it is possible that some of the child or adolescent's characteristics may exert their influences and produce biased results. In particular, behaviors that produce downwardly biased results, such as those associated with anxiety, fatigue, inattention, extreme disappointment, and noncompliance, should be monitored closely.

Commonplace Practices

Two practices for recording and summarizing test session behaviors have become commonplace. The first practice is writing descriptions of behaviors of interest in the margins of the test record. These narrative "behavior observations" may be used to enhance memory when examiners are summarizing test results and writing reports, as well as to contribute to a better understanding of low norm-referenced scores yielded after testing has been completed. These behavior observations are often written concisely. Examples include such notes as "Seems anxious: fidgety, picks at skin around fingernails," "Says hates math," or "Is a live wire: bouncing in chair, grabbing testing materials, out of seat." Although some of these behaviors may be a reaction to test content (e.g., the one about hating math), these behaviors sometimes represent our failures to eliminate or optimally control for construct-irrelevant influences. Because behavior in one-on-one testing environment does not accurately represent behaviors in other environments, you should be more concerned about producing valid test scores than about observing the child or adolescent's reac-

tion to the testing session. You should not sit passively and record behaviors that confound your results; you should intervene. For example, with an adolescent who appears to lack motivation, you could ask, "Is everything OK? You are looking a bit tired," problem-solve based on the response, and terminate the testing session if needed (Kamphaus, 2001). The second practice is summarizing (after testing is complete) noteworthy behaviors displayed during the testing session. Almost every intelligence test record includes a brief checklist or prompts addressing language usage, attention, activity level, attitude, cooperativeness, mood, and self-confidence. Both methods may assist you in describing the testing environment in your report after detailed memories of the session have decayed.

Innovative Practices

Two innovative practices cast light on our understanding of test session behaviors and their potential influences on the validity of assessment. First, A. S. Kaufman, N. L. Kaufman, and their collaborators (see Kaufman & Kaufman, 2004a) have developed a brilliant way to monitor and record potential construct-irrelevant influences on item-level performance: having the examiner review and complete a brief checklist of common influences associated with each subtest. Their *qualitative indicators* include *disrupting indicators*, including not monitoring accuracy, failing to sustain attention, impulsively responding incorrectly, perseverating despite feedback, refusing to engage in a task, being reluctant to commit to a response, being reluctant to respond when uncertain, and worrying about time limits. Conversely, they include *enhancing indicators*, including closing eyes to concentrate, asking for repetition of items, persevering after initial struggles, trying out options, being unusually focused, verbalizing related knowledge, verbalizing a strategy for recall, and working quickly but carefully. Reviewing a checklist of such items during or after the administration of each subtest allows examiners to identify and document potentially confounding influences. Second, rating scales have been developed to assess behavioral excesses that may interfere with test performance. Two of the most well-developed and psychometrically sound instruments of this kind are the Guide to Assessment of Test Session Behavior (Glutting & Oakland, 1993) and the Achenbach System for Empirically Based Assessment Test Observation Form (TOF; McConaughy & Achenbach, 2004). The completion of postassessment rating scales such as the TOF may be especially useful for students in training, because it sensitizes them to unusual and perhaps meaningful behaviors during the testing session. For discussion of these methods, see McConaughy (2005).

Meaningfulness of Behavior Observations

Examiners' narrative descriptions and ratings of observations completed after the testing session are likely to be negatively affected by errors in memory (Lilienfeld, Ammirati, & David, 2012; Watkins, 2009), so we encourage recording behaviors of interest immediately after they occur during testing. The most accurate summaries of behaviors during the test session would probably stem from review of such notes recorded in the protocol. There is no doubt that these varying types of behavior observations can be completed in concert, but we encourage relying first on direct observation and immediate recording of them. We should perhaps conclude, upon reflection, that specific behaviors were not undermining test scores unless notes indicated such throughout protocol.

Posttesting Considerations

Debriefing

We admit to frequently making the mistake of not conducting some sort of debriefing after completing a test session (and not teaching our students to do so!). After a full testing session, it is tempting to say, "We're done today. Thanks for working so hard with me. Let's go back to your classroom," or the like, but it is an ideal practice to spend a bit more time debriefing. First, it may help to maintain rapport with the child or adolescent—especially if multiple testing sessions will be completed. Kamphaus (2001) has suggested honestly acknowledging challenges the child or adolescent faced during testing. Statements like "I realize that you were a little unhappy about being here today, but you tried hard to do what I asked and I really appreciate that" (Kamphaus, 2001, p. 111) may go a long way in building and maintaining your relationship with the child or adolescent. In addition, it may allow you to answer questions or allay fears the child or adolescent may have. Second, if you inquire about the child or adolescent's experience of the test, as well as what he or she liked and did not like, you may identify probable construct-irrelevant influences on test results. Third, debriefing is consistent with the goal of keeping the child or adolescent informed about the outcomes of the testing. Sattler (2008) has suggested repeating some of your introductory statements regarding these outcomes, such as your sharing that you will summarize the results in a report and meet with parents and teachers to discuss the results so that the examinee's experience at school will be improved. We strongly encourage taking additional time at the end of each testing session to debrief the child or adolescent.

Validity of Results

After testing is complete, you must make two decisions. One decision is to determine whether the results are "valid." This validity is not exactly the same type as that discussed in Chapter 5, but the concepts are similar. Essentially, you need to judge whether the intelligence test subtests were administered in a way—and the child or adolescent reacted to them in a way—that permits you to say with confidence that the resulting scores well represent the ability or abilities you targeted. In other words, you must judge whether the inference made about the child or adolescent's cognitive abilities are reasonably accurate, based on his or her test behavior.

You must judge whether the inference made about the child or adolescent's cognitive abilities are reasonably accurate, based on his or her test behavior.

Although statements about the validity of testing results are often made at a global level when printed in reports (see Chapter 8 for examples), we encourage consideration of the meaningfulness of subtest scores, which are the foundation of all other scores for intelligence tests. It is possible that the results from one subtest are invalid and all other results are valid. We encourage you to specify which scores are likely to be invalid and describe what led you to conclude this. Perhaps there is a continuum of validity of assessment. On one end is an ideal testing environment and a perfect standardized administration. There were no interruptions; the child or adolescent was optimally motivated to respond and displayed no behavior problems; and there was consistent performance across items, with no apparent guessing or failure to answer the easiest items. Such sessions are

not uncommon. On the other end is a testing session that goes awry. You are testing in a broom closet near a noisy hallway and a train track; you are not at your best due to caffeine withdrawal; you fumble through the testing materials; and the child or adolescent is unruly and unmotivated. In cases such as the latter, you may conclude that all of the data collected as part of the assessment are meaningless in representing the child or adolescent's cognitive abilities. If so, you should not report the scores in your report or discuss them. You should promptly schedule another testing session and develop a plan for improving it. Furthermore, you should briefly summarize this event in your report, justify your decision not to report your initial results, and describe your rationale for administrating a second intelligence test.

The most difficult validity decisions are those between the extremes on the continuum—those cases in which there was probably some interference in measurement during testing. We have found that these instances are rarely addressed by examiners and included in reports, but we urge giving more attention to them and their influence on test scores. They may stem from influences generally associated with the testing environment. For example, some children and adolescents may be less engaged in the testing late in the session, and the results are lower scores on the last few subtests administered. Others may struggle in completing items from one subtest and seem to be dejected afterward. Still others will react in an unexpected way to standard test directions, test items, and prompts and queries. For example, some will simply not understand the requirements of the task as a whole, despite completing practice items satisfactorily. Others may adopt the strategy of providing you more information than you required—going on and on when defining words or providing the correct answer and also five variants of it. Still others will offer a new response—effectively changing their answer—when queried for more information. These external influences and idiosyncratic reactions almost certainly have an impact on test scores, but it is admittedly difficult to determine when they reflect true ability deficits versus construct-irrelevant influences.

Testing of Limits and Follow-Up Assessment

There are at least two methods for addressing potential construct-irrelevant influences and the validity of test results: (1) testing of limits and (2) follow-up testing using similar measures. *Testing of limits* refers to readministration of subtest items after completion of a test, during which additional cues are given, modeling of correct responses is completed, items are presented in a different fashion, responses are allowed in another modality (e.g., spoken vs. written), time limits are eliminated, and probing questions are asked. The goal is to test hypotheses about why a child or adolescent missed an item or series of items. For example, in cases in which the child or adolescent quickly missed all items on a subtest requiring him or her to repeat orally presented numbers backward, you could use concrete objects (e.g., tiles or blocks with numbers on them) to demonstrate how to reverse numbers presented orally to ensure that the child or adolescent understands what is required. Alternately, you could request that they teach you how to complete an easy item. Afterward, you could administer the items again to determine whether the intervention has yielded dividends. Regardless of the outcome, you should never change raw scores obtained from a standardized test administration, but you should consider explaining the reason for a low score and eliminating the affected score from your report.

Testing of limits has been formalized in some published tests, such as the Wechsler Intelligence Scale for Children—Fourth Edition Integrated (WISC-IV Integrated; Kaplan et al., 2004). It contains 16 "process tests" that reflect changes in the task requirements of subtests through

adaptations to their stimuli, to response requirements, and to the nature of the subtests themselves. Some of these adaptations are minor, such as laying a grid over an image to facilitate reconstruction of that design with blocks; others are substantial, such as requiring children to point to images of numbers in sequence rather than repeat orally presented numbers. Performance on most of these process tests yields norm-referenced scores that can be compared to those stemming from the original standardized administration of subtests from the Wechsler scales. Higher scores on the process tests than on the original subtests are thought to indicate that the process targeted by the adaptation must have interfered with performance on the original subtest. For example, an examiner might conclude that a child or adolescent who performed notably better on the subtest requiring recall of number factors, calendar facts, geography, science, and history when provided with multiple-choice options than when provided with no external cue experienced memory retrieval problems that interfered with his or her initial performance on the subtest without the cues. We worry, though, that repeated exposure to the same items during testing of limits has unknown but probably facilitative effects on the responses of the child or adolescent.

In addition, when we develop hypotheses regarding construct-irrelevant influences on test scores, we often engage in some additional testing employing related measures. If we believe that some distraction during testing, some misunderstanding on the child or adolescent's part, or some other influence unduly affected test scores, we usually turn to resources offered as part of the cross-battery approach (Flanagan, Ortiz, & Alfonso, 2013) to find an alternative subtest purported to measure the same specific ability to test our hypothesis. Often we identify alternative subtests that go about measuring the ability in a different way—somewhat like those included in the WISC-IV Integrated (Kaplan et al., 2004), but from another test. For example, if a child scored poorly on a subtest that requires repeating orally presented numbers in reverse, we might follow up with a related subtest that requires him or her to order orally presented numbers and letters in ascending order. Alternately, we might follow up with a subtest that more closely matches the content and response processes of the one in question. Although we are not certain that the extra effort to examine all of our hypotheses in this vein is always warranted, judicious use of this technique may allow you to distinguish more clearly between low scores indicating ability deficits and low scores due to construct-irrelevant influences.

SUMMARY

It is important to have a strong knowledge of the reasons for assessment, typical assessment processes, and potential influences on test performance that can be controlled during testing or acknowledged when interpreting test results. In this chapter, we have offered detailed descriptions of these processes and equipped you with tools and resources to facilitate and evaluate the validity of your test scores for each child or adolescent you assess.

FORM 4.1

Screening Tool for Assessment (STA)—Parent and Caregiver Screening Form

Child's name: _____ **Glasses/contact lenses: Y/N**

Parent or caregiver's name: _____ **Hearing aid/cochlear implant: Y/N**

Please read each item carefully and circle Yes or No.

1	My child closes or covers one eye when looking at some things.			Yes	No
2	My child frequently squints his or her eyes.			Yes	No
3	My child complains that some images are blurry or hard to see.			Yes	No
4	My child holds objects unusually close to his or her face when looking at them.			Yes	No
5	My child seems to blink a lot.			Yes	No
6	My child becomes frustrated or upset when doing close-up work such as reading, math, and puzzles.			Yes	No
7	My child uses the wrong color names for objects, such as saying that there are purple leaves on trees.			Yes	No
8	My child has a short attention span when coloring or drawing with colors or colored markers.			Yes	No
9	My child has difficulty identifying red or green.			Yes	No
10	My child has difficulty reading when the words are on colored pages.			Yes	No
11	My child does not respond to loud noises sometimes.			Yes	No
12	My child's listening skills are behind what I expect.			Yes	No
13	My child turns up the volume too loud on electronic equipment.			Yes	No
14	My child does not follow spoken directions well.			Yes	No
15	My child often says, "Huh?" or asks you to repeat something you have said.			Yes	No
16	My child does not respond when called.			Yes	No
17	My child's speech is behind what I expect.			Yes	No
18	My child has difficulty pronouncing some words.			Yes	No
19	My child substitutes sounds in words (e.g., *wed* for *red*).			Yes	No
20	My child leaves sounds out of words (e.g., *root* for *fruit*).			Yes	No
21	How much of your child's speech do you understand?	25%	50%	75%	100%
22	How much of your child's speech would a stranger understand?	25%	50%	75%	100%

(continued)

23	My child has difficulty stacking blocks.	Yes	No
24	My child has difficulty fitting puzzle pieces together.	Yes	No
25	My child has difficulty drawing a straight line.	Yes	No
26	My child has difficulty drawing a circle.	Yes	No
27	My child has difficulty printing his or her first name.	Yes	No
28	I can usually read my child's handwriting.	Yes	No
29	A stranger would usually be able to read my child's handwriting.	Yes	No
30	My child is often defiant.	Yes	No
31	My child often refuses to do what I ask.	Yes	No
32	My child breaks a lot of rules at home.	Yes	No

Are there any other issues that might affect how well your child performs during the upcoming testing sessions? Please describe them below.

FORM 4.2

Screening Tool for Assessment (STA)—Teacher Screening Form

Student's name: _____ **Glasses/contact lenses:** **Y/N**

Teacher's name: _____ **Hearing aid/cochlear implant:** **Y/N**

Please read each item carefully and circle Yes or No.

1	The student closes or covers one eye when looking at some things.				Yes	No
2	The student frequently squints his or her eyes.				Yes	No
3	The student complains that some images are blurry or hard to see.				Yes	No
4	The student holds objects unusually close to his or her face when looking at them.				Yes	No
5	The student seems to blink a lot.				Yes	No
6	The student becomes frustrated or upset when doing close-up work such as reading, math, and puzzles.				Yes	No
7	The student uses the wrong color names for objects, such as saying that there are purple leaves on trees.				Yes	No
8	The student has a short attention span when coloring or drawing with colors or colored markers.				Yes	No
9	The student has difficulty identifying red or green.				Yes	No
10	The student has difficulty reading when the words are on colored pages.				Yes	No
11	The student does not respond to loud noises sometimes.				Yes	No
12	The student's listening skills are behind what I expect.				Yes	No
13	The student turns up the volume too loud on electronic equipment.				Yes	No
14	The student does not follow spoken directions well.				Yes	No
15	The student often says, "Huh?" or asks you to repeat something you have said.				Yes	No
16	The student does not respond when called.				Yes	No
17	The student's speech is behind what I expect.				Yes	No
18	The student has difficulty pronouncing some words.				Yes	No
19	The student substitutes sounds in words (e.g., *wed* for *red*).				Yes	No
20	The student leaves sounds out of words (e.g., *root* for *fruit*).				Yes	No
21	How much of the student's speech do you understand?	25%	50%	75%	100%	
22	How much of the student's speech would a stranger understand?	25%	50%	75%	100%	

(continued)

23	The student has difficulty stacking blocks.	Yes	No
24	The student has difficulty fitting puzzle pieces together.	Yes	No
25	The student has difficulty drawing a straight line.	Yes	No
26	The student has difficulty drawing a circle.	Yes	No
27	The student has difficulty printing his or her first name.	Yes	No
28	I can usually read the student's handwriting.	Yes	No
29	A stranger would usually be able to read the student's handwriting.	Yes	No
30	The student is often defiant.	Yes	No
31	The student often refuses to do what I ask.	Yes	No
32	The student breaks a lot of rules at school.	Yes	No

Are there any other issues that might affect how well the student performs during the upcoming testing sessions? Please describe them below.

FORM 4.3

Screening Tool for Assessment (STA)—Direct Screening Test Record

Student's name: _____ **Glasses/contact lenses:** **Y/N**

Date: _____ **Hearing aid/cochlear implant:** **Y/N**

GENERAL SCREENING QUESTIONS

I want to ask you a few questions before we start today.

How are you feeling today?

Did you sleep well last night?

How confident (or nervous) are you feeling about our work together today?

How ready are you to do your best on every task today?
***or* How motivated are you to do well on these tasks today?**

Do you wear glasses or contacts or have trouble seeing?

Do you have any trouble telling colors apart?
***or* Are you color-blind?**

Do you wear a hearing aid or have trouble hearing?

Do you have any trouble writing with a pencil?
***or* How neat is your handwriting?**

Do you have any trouble picking up and moving items such as coins?

Note. Paraphrase and follow up with more specific questions if needed. Use the child or adolescent's oral language during interview to estimate intelligibility of speech.

(continued)

VISION SCREENING

Present Vision Screening items from the STA Direct Screening Response Form.

Look at these letters on this sheet of paper. Please tell me the letters in the top row.

Stimuli	Number incorrect	Total correct
E O P Z T L C D F		/9

Correct errors if necessary, and repeat until all items are correct in sequence.

Without picking up the sheet, tell me the letters in the other rows. There is no reason to go fast. Slow and careful reading is best.

Mark those items missed.

Stimuli	Number incorrect	Total correct
T D P C F Z O E L		/9
D Z E L C F O T P		/9
F E P C T L O Z D		/9

COLOR BLIND SCREENING

Present Color items from the STA Direct Screening Response Form.

See these squares. Name these colors for me. (Alternatively, **What is this color?**)

Stimuli	Correct response	
Black	Yes	No
Blue	Yes	No
Red	Yes	No
Green	Yes	No
Yellow	Yes	No
Purple	Yes	No

Total correct: /6

FINE MOTOR SCREENING

Present items from the Fine Motor page of the STA Direct Screening Response Form. Administer items in sequence as age and developmental level dictates.

If the examinee attempts to rotate the paper, say, **Don't move the paper; keep it in one place.**

1. **Take this pencil and trace this dotted line. Follow on the line, and make one smooth line.**
2. **Now trace this star. Follow on the lines.**
3. **Now trace this circle. Follow on the line.**
4. **Write this sentence, "The dog ran," on this line.**

(continued)

Stimuli	Successful	
Line	Yes	No
Star	Yes	No

Stimuli	Successful	
Circle	Yes	No
Sentence	Yes	No

Total Yes:	/4

HEARING SCREENING

Part A: **Let's play a game about listening. I want you to point to parts of your body. Ready? Listen carefully. I will tell you where to point.**

Administer the initial items slightly louder than normal.

Item	Correct response	
Point to your nose.	Yes	No
Point to your hair.	Yes	No
Point to your face.	Yes	No
Point to your ear.	Yes	No
Point to your chin.	Yes	No
Point to your shoulder.	Yes	No
Point to your mouth.	Yes	No

Total correct, Part A: _____

Part B: **Now I want you to point to the same parts of your body, but this time I want you to close your eyes while I tell you where to point. Ready? Listen carefully.**

Administer these items at a low volume that is above that of a whisper, but still able to be heard by the examinee. Put your hand a few inches in front of your mouth during administration.

Item	Correct response	
Point to your face.	Yes	No
Point to your nose.	Yes	No
Point to your ear.	Yes	No
Point to your hair.	Yes	No
Point to your shoulder.	Yes	No
Point to your mouth.	Yes	No
Point to your chin.	Yes	No

Total correct Part B: _____

(continued)

ARTICULATION SCREENING

Please close your eyes, listen carefully, and say what I say.

Carefully articulate each word at a normal volume. Carefully watch the lips of the child or adolescent while he or she is responding.

If you cannot determine how to score a response, say "Please, say (item) again" as often as needed.

Circle the phonemes in the Sounds column that were incorrectly pronounced. Subtract number of incorrect phonemes from 23 to determine the percentage correct, using the table provided.

Item	Sounds	
Lizard	/l/	/z/
This	*th*	/s/
Wreath	/r/	*th*
Shoes	*sh*	/z/
Seal	/s/	/l/
Think	*th*	—
Mouthwash	*th*	*sh*
Garage	/r/	*zh*
Feather	*th*	/r/
Pillow	/l/	
Zebra	/z/	
Soothe	*th*	
Television	*zh*	
Listen	/s/	
Washer	*sh*	

Total sounds: **23**

Number of incorrect sounds: _____

Number of correct sounds: _____

#Correct	%Correct
0	0%
1	4%
2	9%
3	13%
4	17%
5	22%
6	26%
7	30%
8	35%
9	39%
10	43%
11	48%
12	52%
13	57%
14	61%
15	65%
16	70%
17	74%
18	78%

#Correct	%Correct
19	83%
20	87%
21	91%
22	96%
23	100%

INTELLIGIBILITY

After completing the STA Direct Screening Test, rate the proportion of the child or adolescent's speech you were able to understand.

How much of the child or adolescent's speech did you understand?	Some 25%	Most 50%	Almost all 75%	All 100%

FORM 4.4

Screening Tool for Assessment (STA)—Direct Screening Response Form

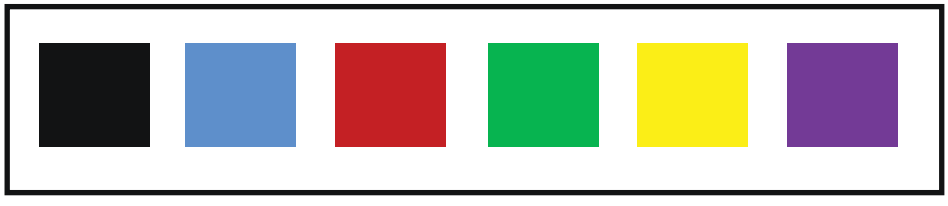
Examinee's name: _____ **Date:** _____

E O P Z T L C D F

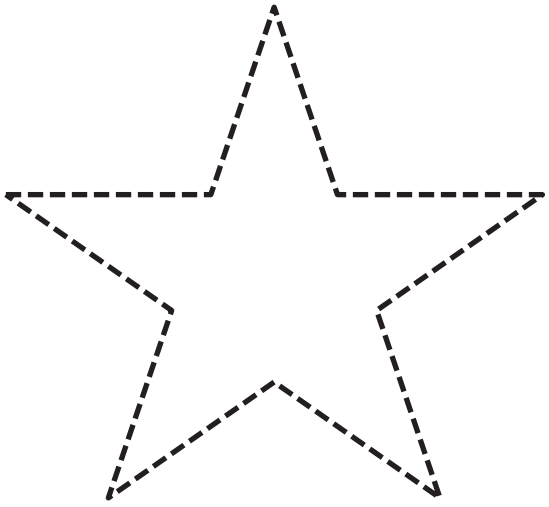
T D P C F Z O E L

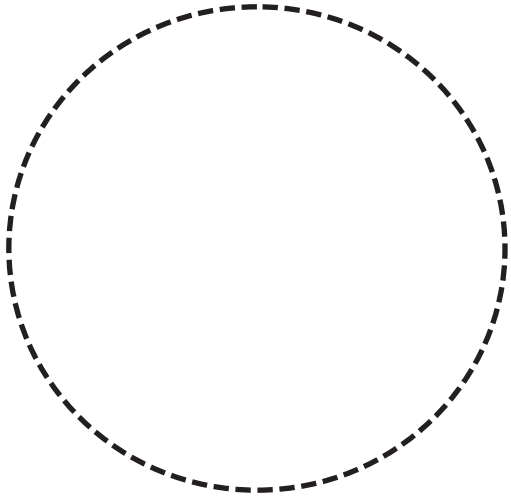
D Z E L C F O T P

F E P C T L O Z D



(continued)





The dog ran.

CHAPTER 5

Selecting the Best Intelligence Tests

Before you can begin to make sense of the meaningful scores from intelligence tests, you should ensure that you fully understand the measurement properties of the tests you are using and their resultant scores. With this information, you can select the best intelligence test for your needs, based in part on the age, ability level, and backgrounds of the clients you serve. This chapter highlights standards guiding the selection and use of tests, and it continues with a review of the most critical characteristics of tests—including norming and item scaling, as well as the reliability and validity of scores.

THE JOINT TEST STANDARDS

In addition to the American Psychological Association's (APA's) Ethical Principles of Psychologists and Code of Conduct and the National Association of School Psychologists' (NASP's) Principles for Professional Ethics (APA, 20002, 2010a; NASP, 2010a; see Chapter 3), you should be familiar with the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 1999). This document was last published in 1999, but at the time of our writing this chapter, a new edition was slated for publication in late 2013. We briefly describe its organization and key components, and then discuss how these components can be applied to selecting the best test for your assessment needs.

The *Standards* document was developed to facilitate scientifically sound and ethical testing practices, as well as to promote evaluation of such practices, and it has traditionally been organized into three sections. The first section addresses test construction, evaluation, and documentation, and its chapters have targeted validity; reliability and errors of measurement; test development and revision; scales, norms, and score comparability; test administration, scoring, and reporting; and sup-

The *Standards* document was developed to facilitate scientifically sound and ethical testing practices, as well as to promote evaluation of such practices.

porting documentation. The second section addresses fairness in testing, and its chapters have targeted the rights and responsibilities of test takers, as well as testing of individuals with diverse linguistic backgrounds and disabilities. The third section addresses testing applications, and its chapters have targeted responsibilities of test users; psychological and educational assessment; testing in employment and credentialing; and testing in program evaluation and public policy.

In this chapter, we focus on general standards and guidelines for test use and general guidelines for psychological testing. Chapter 4 has addressed issues in the assessment process; Chapters 6 and 8 address interpretation and reporting of scores; and Chapter 13 highlights issue related to fairness and nondiscriminatory assessment. As conveyed in the *Standards* (AERA et al., 1999), evaluation of the acceptability of tests and their measurement characteristics involves, at least partially, professional judgment and knowledge of the content domain area. Using the general standards and guidelines for test use from the *Standards* document, we present criteria for evaluating the quality of norming, item scaling, and reliability and validity of scores yielded by intelligence tests.

EXPECTATIONS AND IDEALS FOR NORMING

Norm Samples

The foundation for the scientific study of intelligence is the natural phenomenon of individual differences in cognitive abilities; individuals of all ages differ in their level and speed of performance on cognitive tasks. In the practice of psychology, an understanding of these individual differences is derived from consideration of where an individual's test performance falls when compared to expectations established via a *norm sample*. A norm sample is ideally a large sample of persons whose characteristics are similar to those of the individual being tested (e.g., age); this sample produces results (i.e., *norms*) to which the individual's results can be compared. These norms allow for a better understanding of typical levels of performance (and variation around it) at different levels of development. Ideally, norm samples would include every individual in a population, but doing so is unfeasible. Instead, norm samples should be as large and representative of the targeted population as possible.

Sometimes the term *standardization sample* is used to refer to the norm sample. For the most part, these terms are used interchangeably, but *standardization* refers to the use of the same directions, sample items, practice trials, feedback, and scoring guidelines for everyone completing a test. Standardization is a necessity for accurate and comparable results to be obtained across administration of tests, and without it, accurate norms would be impossible. We suppose that criterion-referenced tests, which typically make no reference to the performance of comparable others, could stem from standardization samples and not norm samples; for most tests, however, standardization samples and norm samples are the same.

Evaluating Norm Samples

Norms affect the accuracy and meaningfulness of the most important data that you will derive from intelligence tests: their norm-referenced scores (see Chapter 6). Accordingly, it is important to scrutinize these tests' norm samples. Typically, norm samples are evaluated according to their recency, their representativeness, and their size. This information is typically provided in techni-

cal manuals accompanying tests, but you should also review the age ranges associated with specific norm groups—those forming the norms and presented in norm tables—to evaluate the appropriateness of norm samples.

Recency

The recency of norms can be evaluated based how near in time the data were collected. As described in the ethical standards discussed in Chapter 3, you should generally avoid using tests that are old and outdated. This recommendation is in place because outdated norms may no longer match the experiences and ability levels of those being tested. For one thing, older tests might include item content that is no longer valid for modern-day test takers, and this could produce biased scores. Furthermore, the evidence indicates rising ability levels across time in populations (called the *Flynn effect*; Flynn, 1984, 1987, 1999, 2007), with older norms producing inflated scores compared to more recent norms. Flynn has estimated that IQ scores increase an average of 3 points per decade, or 0.3 points per year. The increase in IQ implies that scores on older tests are likely to be inflated in comparison to those on newer tests with contemporary norms. Therefore, it is essential to use the most recently normed tests.

To evaluate the recency of norms, you should examine not only the date of a test's publication, but also, and more importantly, the period during which its norming data were collected. You should ideally select intelligence tests that have been normed within the past 10 years. Tests with norming data collected earlier should be used sparingly and should probably be avoided.

Representativeness

The representativeness of norms is equally important. It can be evaluated according to the extensiveness of sampling (1) across political regions (e.g., U.S. states) and across data collection sites; and (2) across demographic characteristics (i.e., gender, race, socioeconomic status, and community size). Of course, norming samples consisting only of participants from a single state or region would not meet the requirement of national norming and would fail to represent the targeted population. Some norms for intelligence tests stem from statistical weighting of contributions by specified subgroups to enhance the representativeness, but if such weighting is not used, a simple way to determine representativeness of the sample is to examine the match between the specific characteristics of the norm sample and those from the referenced census report used as a benchmark for norming. For example, comparisons could be made according to geographic characteristics (e.g., region of the country, population size, and urban vs. rural settings) and demographic characteristics (e.g., race, ethnicity, gender, and socioeconomic status). Discrepancies between specific characteristics of the norm sample and those from the respective census that are greater than 5 percentage points have been reported as indicating oversampling or undersampling (Floyd & Bose, 2003; Merrell, 2008). As discussed in Chapter 13, the absence of a particular individual's demographic group from the norming sample is *not* ipso facto evidence that a test is biased. The test may be biased, but it may not be. Evidence of bias can be determined only by empirical research.

It has been claimed that individuals with disabilities and those who receive extremely high and extremely low scores should be included in the norming samples—and some intelligence test samples make a point to ensure that such cases are included—but it is becoming increasingly common for *extrapolated norms* to be used. Extrapolated norms are derived from statistical manipula-

tions of the distributions of test scores to ensure that they fully represent the population as a whole; they compensate for the fact that the norming data may not, in fact, represent the population well. For example, extrapolated norms can produce an IQ of 185 for children age 5 years old, although no one on the normative sample actually obtained an IQ that high. Fully representative norms based on expansive ability sampling are extremely difficult to obtain, because it is so hard to find sufficient numbers of people with extremely high IQs, and so hard to test those with extremely low IQs due to potentially confounding factors.

Developmental Sensitivity

The age range covered by each segment of the norms must be considered to ensure that they are sensitive to developmental differences, especially during the periods when cognitive growth is most rapid (i.e., up until about 16–20 years of age; see Chapter 2). That is, divisions of the norms should help to disentangle the effects of maturation and experience that are associated with age (as confounds) from individual differences in the display of the targeted abilities (as targeted by norm-referenced scores). Without an effort to eliminate these confounds, the norms would underestimate the abilities of the youngest children at the targeted age level (producing lower scores) and overestimate the abilities of the oldest children in the targeted age level (producing higher scores; Flanagan, Ortiz, Alfonso, & Mascolo, 2006).

Although parents, teachers, and other professionals tend to ground their developmental expectations in a child or adolescent's whole age in years or whole grade levels in school, norm divisions representing narrower periods of time (e.g., a half year, a quarter, or a few months) are often necessary to produce the most accurate norm-referenced scores. Many tests (especially those targeting very young children) include norm level sample blocks representing rather narrow periods of time (e.g., 2 months and 3 months). In contrast, when developmental differences across age groups would not be anticipated (e.g., during the decade of the 30s), norm sample blocks are often much broader (e.g., 5 to 10 years; Flanagan et al., 2006). According to Bracken (2000), in addition to considering the breadth of these blocks, it is important to evaluate the actual norm tables to judge their adequacy—especially when a child is "on the very upper cusp of one age level and who is about to 'graduate' to the next age level" (p. 42). Bracken has shared that one way to evaluate this *sensitivity* of the norms is to review the norm tables and "examine the difference in standard scores associated with a given raw score as you progress from one table to the next. If the standard score increases by large amounts (e.g., $+1\frac{1}{3}$ standard deviations), the test may provide too gross an estimate of ability to instill much confidence in the resultant score" (p. 42). To address this issue, standard score values should be compared at levels near the mean, about one standard deviation above the mean, and about one standard deviation below the mean, beginning with the block associated with the age of the prospective examinee.

In recent decades, many test authors have addressed the developmental sensitivity of norms by using *continuous norming* procedures (see tables in Chapter 7). Rather than deriving norm-referenced scores from descriptive statistics based on individuals in large segments of the norms (e.g., children the same year in age), as previously indicated, sophisticated statistical methods (e.g., curve smoothing) are used to consider the performance across much smaller age segments and to integrate this information across ages, so that a picture of the developmental expectations within the larger grouping are well represented. Test norms developed by using these methods are noteworthy and superior to those developed via more traditional methods.

Norm Sample Size and Size of Norm Blocks

In general, larger norm samples indicate higher-quality norming and more trustworthy norm-referenced scores. It is important to consider the size of each age-based interval of the norms. According to common standards (e.g., Emmons & Alfonso, 2005; Flanagan & Alfonso, 1995; cf. Hammill, Brown, & Bryant, 1992), each 1-year age-based interval of the norms can be judged to be acceptable if it includes at least 100 individuals. This standard seems like a reasonable one to us. However, evaluating this characteristic is sometimes tricky, because some tests present norm sample sizes across wider age ranges (e.g., 1-year increments) in the body of the technical manuals, but rely on norm sample blocks representing narrower time segments (e.g., 3- to 6-month intervals), as previously discussed. In such cases, the mean number of participants per norm block can be estimated by dividing the number of participants reported across the wider age range by the number of norm-related segments by which it is divided. For instance, if 120 children age 4 were reported to compose a 1-year norm block, but norms were calculated in three 4-month blocks (4:0–4:3, 4:4–4:7, and 4:8–4:11), an average of 40 children per norm block would be assumed. This value would be unacceptable based on the goal of including at least 100 children in this group, and we believe that an acceptable absolute low-end standard should be 30 children per norm block.

In general, larger norm samples indicate higher-quality norming and more trustworthy norm-referenced scores.

SCALING

Range of Norm-Referenced Scores

Because you may be called on to assess children or adolescents suspected of having intellectual disability (ID; see Chapter 10) or intellectual giftedness (see Chapter 11), it is important that you know the range of scores provided by the intelligence tests you are evaluating. In addition, in order to differentiate between ability levels at every point across this range, you should consider item gradients for intelligence test subtests.

Scale Floors

A *scale floor* refers to the lowest norm-referenced score that can be obtained. When these floors are too high, the full range of ability at its lowest levels cannot be assessed, because, at the subtest level, items are too difficult for individuals with low ability; there are too few "easy" items for them. As a result of such problems, even when a child performs poorly on a subtest, answering only one item (or no items) correctly, the child will earn a score that is relatively close to what is considered average. Subtest scores with insufficient floors will overestimate the abilities of those who are near the lowest end of the ability range, and this overestimation will be transfered to all of the composite scores to which that subtest contributes.

Floors for subtests can be evaluated by examining norm tables and using score software; if a raw score (i.e., the sum of item scores) of 1 does not yield a norm-referenced score equal to or exceeding two standard deviations below the normative mean (a deviation IQ score of 70 or lower, a *T* score of 30 or lower, or a scaled score of 4 or lower; see Chapter 6), a *floor violation* is

evident (Bracken, 1987; Bracken, Keith, & Walker, 1998; Bradley-Johnson & Durmusoglu, 2005). For composite scores that stem from summing norm-referenced scores from subtests, floor violations are apparent if the lowest basic score (e.g., the sum of scaled scores) is associated with a norm-referenced score less than two standard deviations below the mean. Floor violations are most frequently apparent at the youngest age levels targeted by intelligence tests.

Scale Ceilings

A *scale ceiling* refers to the highest norm-referenced score that can be obtained. When ceilings are too low, the full range of ability at its highest levels cannot be assessed. This scenario is the opposite of the one occurring when children or adolescents with low ability are assessed. As a result of such problems with ceilings, those with high ability are likely to have their abilities underestimated. Even when a child performs extremely well on a subtest, answering every item correctly, the child will earn a score that is relatively close to what is considered average. There are too few "difficult" items on the subtest to sufficiently challenge such children at this level.

Ceilings for subtests can also be evaluated by examining norm tables and using score software; if the highest possible raw score does not yield a norm-referenced score equal to or exceeding two standard deviations above the normative mean (i.e., a deviation IQ score of 130 or higher, a *T* score of 70 or higher, or a scaled score of 16 or higher; see Chapter 6), a *ceiling violation* is evident (Bracken, 1987; Bracken et al., 1998; Bradley-Johnson & Durmusoglu, 2005). For composite scores, ceiling violations can be identified if the highest basic score (e.g., the sum of scaled scores) is associated with a norm-referenced score less than two standard deviations above the mean. Ceiling violations are most frequently identified at the oldest age levels targeted by intelligence tests.

Item Scaling

Test authors typically devote much time and effort to developing, selecting, and scaling items for their instruments. For most subtests on intelligence tests, items are selected and scaled from easiest to most difficult. Examinees begin with items that most individuals of the same age would successfully complete, progress through items that are neither too easy nor too difficult for them, and reach items that are on the threshold of their current knowledge and skills. In order to evaluate the quality of an intelligence test, you should consider the methods used to accomplish such scaling and should evaluate evidence suggesting effective scaling.

Scaling Techniques

Traditional techniques used for item scaling include *item difficulty analysis*, in which items are evaluated according to the percentage of examinees passing those items. After administering the preliminary set of items to large groups of individuals, test authors using these techniques can identify patterns of scores across items and then place the items in order from those passed by the most individuals to those passed by the least individuals. During the past 30 or 40 years, *item response theory* (often referred to as only IRT, although we continue to use the full term) techniques have replaced more traditional techniques like item difficulty analysis. Item response theory analysis considers not only item difficulty, but also the relation between passing the item and the sum of all the items considered in the analysis (often called *item discrimination*) and ran-

dom error (often called *guessing*). One common type of item response theory analysis is referred to as *Rasch modeling* (Rasch, 1960). This analysis, focused on accurately measuring the latent trait (or ability) underlying performance on an item, produces a purer method of evaluating item difficulty than any other single traditional technique, and for most subtests, it is the optimal method for scaling items. When you review technical information about an intelligence test, you should expect authors to have used item response theory to scale items.

Item Gradients

Appropriately scaled items should not only proceed from easy to difficult, but should also do so without rapidly progressing from easy to difficult items. An example of inappropriate item scaling for an achievement test would be a math subtest that includes easy single-digit multiplication items, advanced trigonometry items, and no items of intermediate difficulty. Such inappropriate scaling, as revealed through *item gradient violations*, can also be identified by a careful review of norm tables.

Ideally, there should be a consistent relation between item scores and norm-referenced scores as they both increase; this relation is referred to as the *item gradient*. According to Krasa (2007), "When the item gradient . . . is too steep, it does not sufficiently absorb 'noise'—that is, errors irrelevant to the construct being tested (such as carelessness or distractibility) can lead to an abrupt change in standard score that does not reflect a true difference in the ability being tested" (p. 4). When this standard is not met, huge jumps in ability estimates (as evidenced by the norm-referenced scores) could be the product of guessing correctly on a single item. For example, with woefully inadequate item gradients, an individual could go from having a slightly-above-average norm-referenced score of 105 to a well-above-average norm-referenced score of 115 because he or she guessed correctly on one item. Conversely, failing one additional item would cause the score to drop precipitously.

Item gradient violations for subtests, which are most closely linked to item-level performance, can be identified if there is not at least one raw score point associated with each one-third of a standard deviation unit in the norms (Bracken, 1987). For example, deviation IQ scores should not change more than 5 points per 1 raw score point change; *T* scores should not change more than 3 points per 1 raw score point change; and scaled scores should not change more than 1 point per 1 raw score point change (see Chapter 6). Even though we do not recommend routine interpretation of norm-referenced scores derived from subtests (see Chapter 6), they should not be ignored, because subtest scores contribute to the more reliable and valid composite scores. However, composite scores, due to their pooling of variability across subtests, should absorb most of the "noise" associated with item gradient violations and other sources of error in measurement.

RELIABILITY

Definition

It is important that the results of your assessment be as precise and consistent as possible. Certainly, none of us would want scoring errors to affect test scores, and we would look askance if we learned that someone obtained an IQ in the Superior range one day and an IQ in the Low Average range the following day after taking the same test. The term *reliability* is used to represent this

valued score characteristic—consistency across replications. Quantitative values targeting reliability represent the extent to which unexplained and apparently random inconsistency across replications, called *measurement error* or *random error*, affects test scores. These unpredictable fluctuations are unavoidable effects of assessment; no test produces a perfect, 100% replicable measurement of any phenomenon. Even in the natural sciences, tools targeting physical measurements are affected by error. For example, rulers and tape measures expand and contract as the temperature fluctuates; even quantum clocks, which are probably the most accurate measures of time, vary slightly (e.g., by 1 second in a billion years) under some conditions.

The term *reliability* is used to represent this valued score characteristic—consistency across replications.

In testing, we most frequently attribute measurement error to the person taking the test, and particularly to the person's variation in determination, anxiety, and alertness from item to item or from day to day. Furthermore, guessing correctly (producing spuriously high scores) comes into play, as do memory retrieval problems (leading to incorrect responses and producing spuriously low scores). In the same vein, those taking the test may respond in different ways to items because of prior experiences. For instance, they may fail items that are typically easier for others (e.g., early presidents of the United States) because they are uninterested in the item content, but may answer items that are typically more difficult for others (e.g., earth science) because they have a special interest in that area. In addition to fluctuation due to the person, external factors may also produce these fluctuations. Examples include the time of day in which the test is taken, distractions in the testing environment, and examiners' deviations in administration and scoring.

Reliability is an important precondition for the validity of tests results and their interpretations. According to the *Standards* (AERA et al., 1999), "To the extent that scores reflect random errors of measurement, their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision making is limited" (p. 31). This relationship between reliability and validity is an important one. In fact, it is this relation that leads us and many other scholars to discourage interpretation of less reliable intelligence test subtest scores and encourage interpretation of more reliable composite scores, such as IQs (see Chapter 6).

Evaluating Reliability and Determining Effects of Error

Although we expect variability stemming from the characteristics of examinees and the circumstances external to them, we assume in our measurement that there is consistency in scores across replications. This consistency in intelligence test scores is typically evaluated by using three methods producing reliability coefficients: *internal consistency*, *test–retest reliability*, and *scorer consistency*. Coefficients close to 1.00 indicate high levels of reliability and acceptable levels of measurement error, whereas coefficients below .70 indicate low reliability and unacceptably high levels of measurement error (Hunsley & Mash, 2008). For scores from intelligence tests and high-stakes diagnosis and eligibility decisions, minimal standards for reliability should be far higher.

Internal Consistency

The type of reliability coefficient most often reported in test technical manuals focuses on item-level consistency evident in a single testing session. These internal consistency coefficients tend

to be the highest of all reliability coefficients, and they are most frequently used to calculate confidence interval values (see Chapter 6). They are often reported as *Cronbach's coefficient alpha*, which represents the average item-to-item relations across the entire scale in question. Alternately, split-half reliability analysis, which examines the relations between odd and even items or between items from the first half and second half of a scale, may be employed. If we assume that all intelligence test subtest items are positively correlated, the greater number of items a subtest includes, the higher its internal consistency reliability coefficient will be. This phenomenon is predicted by the *Spearman–Brown prophecy formula* (Brown, 1910; Spearman, 1910). Although these reliability analyses are commonly used to determine the internal consistency of intelligence test subtests, the internal consistency of resulting composites, such as IQs, are often determined by using subtest internal consistency coefficients and subtest intercorrelation values (Nunnally & Bernstein, 1994).

In general, the closer the analysis of reliability is to the item level, the lower the reliability will be. Items are most strongly affected by measurement error, but these effects are diminished as more and more items are considered in concert—from subtest scores, to composite scores formed from only a few subtests, and to composite scores stemming from numerous subtests. Thus, the most global composites are always the most reliable. When high-stakes decisions are to be made, we suggest using a lower-end standard of .90 for internal consistency reliability coefficients (Bracken, 1987; McGrew & Flanagan, 1998); scores with values below .90 should not be interpreted. We recommend that scores with internal consistency reliability coefficients of .95 or higher should be targeted.

Test–Retest Reliability

Evaluation of consistency across replications is most apparent when the same instrument is administered twice across a brief period (typically a month or less). When the relations between scores stemming from these two administrations are quantified by using a statistical technique called the *Pearson product–moment correlation*, an understanding about reliability across time is yielded. To the surprise of some, this technique is an appropriate one, because it is insensitive to differences in the magnitude of scores from the initial testing to the follow-up testing. As a consequence, the resultant correlation value, the *Pearson coefficient*, will reflect only the extent of relative consistency.

Test–retest reliability coefficients are commonly reported for intelligence test subtests and composites, and they may be the only type of reliability coefficient reported for speeded subtests, because traditional internal consistency reliability coefficients cannot be calculated for speeded subtests (unless advanced item response theory techniques are used). Test–retest reliability coefficients are degraded by multiple influences; both the length of the interval between test and retest and the nature of the ability being targeted should be considered in evaluating them. Typically, the longer the interval between initial and follow-up testing (e.g., 1 month vs. 6 months), the lower the reliability coefficient will be. Furthermore, if the ability being targeted tends to vary because of influences associated with the examinee and the effects of testing environment (e.g., Processing Speed), test–retest reliability coefficients will tend to be lower than those for abilities that tend to be more resistant to these influences (e.g., Crystallized Intelligence). Test–retest reliability coefficients tend to be lower than internal consistency reliability coefficients, and when high-stakes decisions are to be made, we suggest using a lower-end standard of .90 (Bracken, 1987; McGrew & Flanagan, 1998).

Scorer Consistency

Those persons scoring intelligence tests also produce measurement error in scores—especially when scores are based on subjective judgments about the quality of responses. This type of measurement error is most frequently quantified by *interrater agreement* indexes and *interrater reliability* coefficients. Interrater agreement indexes stem from analysis of individual items across a subtest, and they are typically reported as a percentage (representing *percentage agreement*) and not as a coefficient. For example, if two raters agree that 5 of 10 responses should earn a point and that 3 of the remaining 5 responses should not earn a point, then their percentage agreement in scoring is 80%. In contrast, interrater reliability coefficients typically stem from analysis of summed item scores across items obtained from independent scoring of responses across two examiners; they are typically the result of Pearson product–moment correlations. To compare and contrast these two indexes of scorer consistency, review Table 5.1. Across the 15 items scored on a 3-point scale by two raters, the scores were almost identical: 17 for Rater 1 and 18 for Rater 2. Furthermore, the average item score for Rater 1 was 1.13, whereas it was 1.20 for Rater 2. When the item-by-item agreement in scoring was considered (see "Agreement" column in Table 5.1), the interrater agreement index (reported as percentage agreement) was 67%, because only 10 of the 15 items were scored exactly the same way. When the correspondence of the rank ordering of item-level scores above and below their respective rater-specific means was considered and reported as a Pearson coefficient, the interrater reliability coefficient of .76 was somewhat modest, but higher in magnitude than the interrater agreement index.

Based on standards for interrater reliability and interrater agreement for assessment instruments targeting child and adolescent behavioral and emotional problems (e.g., Achenbach, McConaughy, & Howell, 1987; Floyd & Bose, 2003; Hunsley & Mash, 2008), interrater reliability levels of .60 or higher and interrater agreement levels of 60% or higher are desirable. Because the range of responses is much narrower, and scoring tends to be clearer with intelligence tests subtests than many other assessment instruments, reasonable lower-end standards for scorer consistency should be .80 for interrater reliability and 80% for interrater agreement.

VALIDITY

Definition

Whereas *reliability* refers to consistency across replications (e.g., across items, across multiple administration of the same tests, or across scorers), *validity* refers to representing the concept or characteristic being targeted by the assessment instrument in a complete and meaningful way. According to the *Standards* (AERA et al., 1999),

> Validity refers to the degree to which evidence and theory support the interpretations of test scores by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. . . . When test scores are used or interpreted in more than one way, each intended interpretation must be validated. (p. 1)

As apparent in this definition, tests do not possess validity; it is a misnomer to refer to a "valid test." Instead, you should refer to "validity of score-based interpretations" and know that that this con-

TABLE 5.1. An Item-Scoring Example Yielding Interrater Agreement and Interrater Reliability Coefficients

			Agreement analysis		Pearson product–moment correlation analysis				
Item number	Rater 1 score	Rater 2 score	Agreement	Agreement score	Rater 1 deviation from *M*	Rater 2 deviation from *M*	Product of deviations from *M*	Squared Rater 1 deviation from *M*	Squared Rater 2 deviation from *M*
1	2	2	Yes	1	0.87	0.80	0.69	0.75	0.64
2	2	2	Yes	1	0.87	0.80	0.69	0.75	0.64
3	2	2	Yes	1	0.87	0.80	0.69	0.75	0.64
4	2	2	Yes	1	0.87	0.80	0.69	0.75	0.64
5	1	2	No	0	–0.13	0.80	–0.11	0.02	0.64
6	2	1	No	0	0.87	–0.20	–0.17	0.75	0.04
7	2	2	Yes	1	0.87	0.80	0.69	0.75	0.64
8	1	1	Yes	1	–0.13	–0.20	0.03	0.02	0.04
9	1	2	No	0	–0.13	0.80	–0.11	0.02	0.64
10	1	1	Yes	1	–0.13	–0.20	0.03	0.02	0.04
11	0	0	Yes	1	–1.13	–1.20	1.36	1.28	1.44
12	1	0	No	0	–0.13	–1.20	0.16	0.02	1.44
13	0	1	No	0	–1.13	–0.20	0.23	1.28	0.04
14	0	0	Yes	1	–1.13	–1.20	1.36	1.28	1.44
15	0	0	Yes	1	–1.13	–1.20	1.36	1.28	1.44
Sum	17	18		10			7.60	9.73	10.40
Average	1.13	1.20							

Percentage agreement

$$\frac{\text{Agreements}}{\text{Total items}} = \frac{10}{15} = \textbf{.67, or 67\%}$$

Pearson correlation

$$r = \frac{\text{Sum of products of deviations for raters}}{\text{Square root of (squared Rater 1 Deviation from } M \text{ times squared Rater 2 deviation from } M)} = \frac{7.60}{\text{Square root of } (9.73 \times 10.40)} = \textbf{.76}$$

cept of validity is conditional, based on the intended uses of those scores. For example, an IQ from a test targeting high-ability preschool students might be valid for identifying intellectual giftedness (see Chapter 11), but if several of its subtests contributing to the IQ demonstrated floor violations, it would not be valid for identifying ID (see Chapter 10). Thus, details and conditions should be specified when you are discussing validity. Asking yourself, "Do I have solid evidence supporting my use of these scores to reach my goals?" addresses this issue well.

***Validity* refers to representing the concept or characteristic being targeted by the assessment instrument in a complete and meaningful way.**

As you consider validity evidence supporting uses and interpretations of intelligence tests scores, you should consider *construct validity* as an overarching conceptual framework. The term *construct* refers to the concept or characteristic that the assessment instrument is intended to measure, and in the case of intelligence tests, the construct is typically a cognitive ability. For example, measures of psychometric *g* represent abstract reasoning and thinking, the capacity to acquire knowledge, and problem-solving ability (Neisser et al., 1996). From this perspective, both test authors and test users should consider and articulate what construct is being targeted by all scores. In doing so, they must consider both (1) how fine-grained or global the intended interpretation is, and (2) what evidence has supported their favored interpretation. For example, does an intelligence test subtest requiring children to provide definitions to English words measure psychometric *g*, Crystallized Intelligence, word knowledge, the ability to articulate word definitions, listening ability, an enriched language environment, expressive language skills, long-term memory, concept formation, or executive system functioning? Of course, it is extremely rare that interpretations of scores are limited to only one meaning, but the wide array of constructs presented in this sample seems to represent what Kelley (1927) called the *jingle–jangle fallacy*. The *jingle fallacy* refers to using the same label to describe different constructs, and the *jangle fallacy*, which is more relevant to this example, refers to using different labels to describe similar constructs. Because of such problems in selecting the targeted construct, you must rely on theory and prior research (as described in Chapters 1 and 2) to develop an explicit statement of the proposed interpretation of your scores or score patterns. This validity argument can then be evaluated on the basis of existing evidence.

Evaluating Validity Evidence

Although the classic tripartite model of validity (focusing on apparently distinct types of validity—*content*, *criterion-related*, and *construct*) is still employed in some test technical manuals and in some of the psychology literature, this model is outdated. Modern conceptions of validity represent validity as a unitary concept; evidence of all types informs the construct validity of the inferences drawn from assessment results and subsequent decisions. Consistent with the *Standards* (AERA et al., 1999), this validity evidence can be compartmentalized into five validity strands: (1) evidence based on test content, (2) evidence based on response processes, (3) evidence based on internal structure, (4) evidence based on relations with other variables, and (5) evidence based on the consequences of testing.

A body of validity evidence should be evaluated by considering potential confounds in measurement that may undermine valid interpretations. The *Standards* document refers to these

potential confounds as rival hypotheses and encourages consideration of both *construct underrepresentation* and *construct-irrelevant variance*. According to the *Standards* (AERA et al., 1999),

> Construct underrepresentation refers to the degree to which a test fails to capture important aspects of the construct. It implies a narrowed meaning of test scores because the test does not adequately sample some types of content, engage some psychological processes, or elicit some ways of responding that are encompassed by the intended construct. Take, for example, a test of reading comprehension intended to measure children's ability to read and interpret stories with understanding. A particular test might underrepresent the intended construct because it does not contain a sufficient variety of reading passages or ignored a common type of reading material. As another example, a test of anxiety might measure only physiological reactions and not emotional, cognitive, or situational components. (p. 10)

Construct irrelevance refers to the degree to which test scores are affected by processes that are extraneous to the test's intended construct. The test scores may be systematically influenced to some extent by components that are not part of the construct. On a reading comprehension test, construct-irrelevant components might include an emotional reaction to the test content, familiarity with the subject matter of the reading passages on the test, or the writing skill needed to compose a response. On an anxiety self-report instrument, a response bias leading to underreporting of anxiety might be a source of construct-irrelevant variance.

Using the concept of construct validity espoused in the *Standards*, and considering both construct underrepresentation and construct irrelevance as contributors to rival hypotheses, you should evaluate evidence for its contribution to interpretations and to revealing potential sources of invalidity. We discuss each type of validity evidence and address its contribution to identifying construct underrepresentation and construct irrelevance in the sections that follow.

Content

Evidence based on *test content* refers to substantiation that an instrument's items accurately represent the targeted construct or constructs in a complete, accurate, and unbiased manner. According to the *Standards* (AERA et al., 1999), "test content refers to the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring" (p. 11). As evident in Table 5.2, item development based on theory, prior literature, and existing diagnostic systems can support the validity of test items. After these items are developed, their evaluation by content experts and after use in item tryouts in the field contribute to such evidence. Although validity evidence based on content is often based on only human judgment, statistical methods may also be used to test validity arguments.

Response Processes

Evidence based on *response processes* refers to substantiation of the real or hypothesized behaviors that test takers follow when completing items. Typically, inferences made about psychological processes or cognitive operations must be drawn from responses to test stimuli. As evident in Table 5.2, these responses can be inferred from review of test stimuli. For example, cognitive tasks can be dissected into their component operations (see Carroll, 1993); reviewers can evaluate test items to determine their match with the targeted response processes; and text can be analyzed

by using readability and cohesion metrics (Flesch, 1949; Graesser, McNamara, Louwerse, & Cai, 2004). Individuals taking the test can inform us about the accuracy of our validity arguments. For example, test takers may be asked to "think aloud" during test completion or to respond to questions about strategies they used (see Ericsson & Simon, 1993), and their responses can be evaluated to identify themes consistent with the targeted constructs. More sophisticated methods, such as eye-tracking technology and recording of response times (see Carpenter, Just, & Shell, 1990), also provide validity evidence based on response processes. Despite the promise of these methods, they are rarely applied to intelligence tests and almost never addressed in test technical manuals.

TABLE 5.2. Definitions of and Methods to Demonstrate the Five Strands of Validity Evidence

Evidence based on test content—substantiation that an instrument's items accurately represent the targeted constructs in a complete, accurate, and unbiased manner.

- Development of items based on strong theory, literature review, established educational or psychiatric diagnostic classifications, and review of case histories
- Expert analysis of gender, racial, cultural, or age bias in items
- Review and item tryouts by test users in applied settings
- Statistical analyses of items (e.g., differential item functioning, point–biserial correlations, and item characteristic curve analyses)

Evidence based on response processes—substantiation of the real or presumed behaviors that test takers exhibit when completing subtest items.

- Evaluation of instrument instructions and response formats
- Observations of test takers' behaviors (e.g., eye movements) during completion of items
- Interviews with test takers about thought processes during completion of subtest items
- "Think-aloud" protocols with test takers during completion of subtest items
- Task decomposition analyses of subtest items

Evidence based on internal structure—substantiation that an instrument's item-level or summative scores are related to other measures from the instrument in the manner expected.

- Correlations between items within a subtest
- Correlations between subtest scores
- Exploratory factor analysis (EFA)
- Confirmatory factor analysis (CFA)

Evidence based on relations with other variables—substantiation that item-level or summative scores from an instrument relate in a systematic way with other measures, such as scores from other instruments, demographic variables (e.g., age and gender), and educational or diagnostic classifications.

- Correlations with measures of the same or similar constructs
- Correlations with measures of distinct or dissimilar constructs
- Correlations with scores from well-validated instruments
- Prediction of current or future phenomena
- Group difference analyses (a.k.a. clinical group comparisons)

Evidence based on the consequences of testing—substantiation that scores and decisions based on them produce intended and not unintended consequences for those completing the test.

- Evaluation of treatment utility
- Evaluation of classification rates of racial/ethnic groups as disabled versus not disabled

Note. Content is based in part on Table 2 from Floyd and Bose (2003).

Internal Structure

Evidence based on internal structure refers to substantiation that a test's items or resultant scores are related to other variables from the test in the expected manner. (If the relations are with some other variables external to the test, another type of validity evidence is considered, as we discuss in the next section.) As evident in Table 5.2, correlations between item scores and subtest or composite scores, between subtest scores, and between composite scores provide such evidence. These correlations are typically Pearson correlations or variants of them. In some instances, correlations between item scores, as also evaluated in internal consistency analysis (described previously), contribute such validity evidence. Perhaps the most sophisticated methods for examining internal structure of tests are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), as described in Chapter 1. These analyses may use item scores to study the latent variables underlying patterns of correlations, but most of the research with intelligence tests (especially that presented in test technical manuals) focuses on relations between subtest scores. As made evident in Chapter 1, factor analysis has provided strong evidence for the structure of human cognitive abilities. It helped to form the foundation for psychometric theories of intelligence, and it continues to offer insights today (Keith & Reynolds, 2010, 2012).

External Relations

To establish the meaning of test scores, they must be related to external criteria. The main goal when investigating the *external relations* of a test is to examine the pattern of relations between measures of the constructs targeted by the test and other measures and variables, such as scores from other instruments, demographic variables (e.g., age and gender), and educational or diagnostic classifications. Construct validity is supported when the pattern of relations between test scores and the external criteria is both rational and consistent with hypotheses based on theory. Conversely, construct validity is not supported when the pattern of relations cannot be explained by hypotheses based on theory (Benson, 1998). This type of validity evidence seems to be most prevalent in test technical manuals and in the research literature; it often composes more than half of the validity evidence presented in test technical manuals.

Evidence of external relations includes correlations between measures of the same or similar constructs, often called *convergent relations*, as well as correlations between measures of dissimilar constructs, often called *discriminant relations*. Analysis of convergent and discriminant relations tends to be theoretically focused, whereas analysis of *criterion-related validity* tends to be more practical. For example, it makes sense that scores from a new intelligence test would correlate highly with those from an older "classic" intelligence test if both were administered to the same children over a brief period. Such evidence would yield *concurrent validity* evidence for the new test when compared to the "gold standard" criteria yielded by the older, classic intelligence test. Furthermore, when examining criterion-related validity, important social and educational outcomes may serve as criteria for comparison. For example, as described in Chapter 1, IQs demonstrate strong *predictive validity* evidence by yielding sizeable, positive correlations with long-term academic outcomes. Again, correlations reflecting convergent and discriminant relations and concurrent and predictive validity evidence are typically Pearson correlations. Finally, validity evidence from external relations can also surface from comparisons of known groups (often

clinical groups, such as children with learning disabilities or attention-deficit/hyperactivity disorder [ADHD]) known to differ on the construct being measured. Such results produce what some researchers (e.g., Floyd, Shaver, & McGrew, 2003; Haynes, Smith, & Hunsley, 2012) call *discriminative validity* (not discriminant validity) evidence.

Consequences

Evidence based on the *consequences* of testing refers to substantiation that scores and decisions based on them produce intended and not unintended consequences for those taking the tests, depending on the purpose of the assessment. In terms of intended consequences, some assert that *treatment utility* (Hayes, Nelson, & Jarrett, 1987; Nelson-Gray, 2003)—evidence that measurable benefits stem from test interpretation—provides positive evidence supporting intended consequences. Others have argued that overrepresentation and underrepresentation of certain racial/ethnic minority groups in special education constitute evidence of invalidity in assessment. This type of validity evidence, however, is understudied and poorly understood (see Braden & Kratochwill, 1997; Cizek, Bowen, & Church, 2010). Many researchers are uncertain about the inclusion of this type of validity evidence along with the other four, because it does not appear to be a relevant criterion for evaluating the validity of certain instruments. For example, do yardsticks provide invalid measures of height because basketball players are taller on average than the general population? Obviously, they are not. On the one hand, we appreciate that the inclusion of this type of validity evidence challenges us to enhance the body of support for the positive outcomes of test results; on the other hand, we fear that excessive emphasis on negative consequences—without carefully designed scientific research studies and clear thinking—will lead to unfounded animosity directed toward testing in schools and related settings, due to presumed invalidity and bias. (See Chapter 13 for further discussion of test bias.)

Making Sense of Validity Evidence

As we have stated previously, when validity evidence is being considered, details and conditions should be specified. Asking, "Do I have solid evidence supporting my use of these scores to reach my goal?" ensures consideration of these details and conditions. According to the *Standards* (AERA et al., 1999), "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (p. 17). Many argue that the validation of any instrument and its scores is an ongoing process, but there is probably a point at which there is a "good enough" body of validity evidence for the provisional application of test score interpretation.

When validity evidence is being considered, details and conditions should be specified.

SELECTING THE BEST TEST FOR YOUR NEEDS

Following the *Standards*, you should be able to (1) consider both the context of the assessment and the personal characteristics and background of each child or adolescent you are slated to assess and (2) select tests that most accurately measure the child or adolescent's cognitive abilities. However,

doing so is no easy feat. You must first consider the stakes involved in the assessment. Most decisions involving intelligence tests are high-stakes decisions, such as diagnosis or eligibility determination. These decisions require higher standards of evidence, and this requirement should lead you to employ a more conservative approach to score interpretation. Then, you must consider the quality of the norms, the reliability of resultant scores, and validity evidence for the intelligence tests available to you. These steps can easily be completed in general, but you must also consider carefully the individual's characteristics, including age, presenting problems, sensory acuity, motor development, language proficiency, background characteristics, and so on, when selecting intelligence tests to administer to him or her.

In order to address these goals, you should seek out published test reviews of intelligence tests. Perhaps the best source for such reviews is the Buros Center for Testing *(www.buros.org)*, formerly the Institute for Mental Measurements, which publishes a series of *Mental Measurement Yearbooks*. Furthermore, several peer-reviewed journals—including *Assessment for Effective Intervention*, the *Canadian Journal of School Psychology*, and the *Journal of Psychoeducational Assessment*—publish test reviews. Although such narrative reviews do not provide systematic, objective evaluations of the tests they target, they can pave the way for your careful review of tests.

The benefit of reading a test's technical manual from cover to cover, however, cannot be understated. As you do, you should pay particular attention to the norming process, review the norm table for scaling problems, evaluate the reliability estimates for the scores you are likely to interpret, and consider the sources of validity evidence. Furthermore, you should consider these test properties and score properties for the specific child or adolescent you will be testing. You should, for example, compare your child's characteristics to those of the norm sample and consider scaling issues subtest by subtest, reliability estimates score by score, and validity evidence appropriate for the child's age and for the referral concern. To assist you in this process, we have reviewed, evaluated, and described characteristics of the most prominent full-length intelligence tests, nonverbal intelligence tests and related composites, and brief and abbreviated intelligence tests in Chapter 7. We have also provided, in Form 5.1, a form for evaluating and selecting the best intelligence test for your needs. This checklist addresses basic demographic characteristics, the referral concern, results from your screening (see Chapter 4), test norming, and measurement properties associated with varying score types.

SUMMARY

This chapter has highlighted standards guiding the selection and use of tests, and it has reviewed their most critical characteristics. We encourage you to promote measurement integrity through careful test reviews and test selection before you begin testing, rather than calling on this information after you have noticed an oddball score or two that does not fit with the remainder of the information from your assessment. The intelligence tests available to you are stronger than ever before. Your efforts to choose the best tests for your needs should yield dividends in your producing more accurate and meaningful results for those children or adolescents you test, as well as for their families and their educators.

FORM 5.1

Intelligence Test Measurement Properties Review Form

Name of child or adolescent: _____

Age in years and months: _____

Referral concern: _____

_____

Screening Results

- ☐ Visual acuity problems
- ☐ Color blindness
- ☐ Auditory acuity problems
- ☐ Articulation problems
- ☐ Fine motor problems
- ☐ Noncompliance
- ☐ Limited English proficiency
- ☐ Acculturation problems

Targeted intelligence test: _____

Norming Information

Last year of norming data collection (see tables in Chapter 7) _____

☐ 10 years ago or less ☐ 15 years ago or less ☐ More than 15 years ago

Age range of norm table block applicable to examinee (e.g., 3.0- to 3.3-year-olds or 7-year-olds):

_____

Number or estimated number of participants in norm table block (see Age unit/block in tables in Chapter 7)

☐ 100 or more ☐ 50 to 99 ☐ 49 to 30 ☐ Under 30

Score Type

Stratum III composites (i.e., IQs)

Maximum norm-referenced score: _____ ☐ Ceiling violation

Minimum norm-referenced score: _____ ☐ Floor violation

Internal consistency reliability: _____ ☐ Under .95

Test–retest reliability: _____ ☐ Under .90

Validity evidence: ☐ Content ☐ Response processes ☐ Internal structure ☐ External relations ☐ Consequences

Stratum III composites (i.e., IQs)

Maximum norm-referenced score: _____ ☐ Ceiling violation

Minimum norm-referenced score: _____ ☐ Floor violation

Internal consistency reliability: _____ ☐ Under .95

Test–retest reliability: _____ ☐ Under .90

Validity evidence: ☐ Content ☐ Response processes ☐ Internal structure ☐ External relations ☐ Consequences

(continued)

Stratum II composites (e.g., broad ability composites)

Score type: ☐ Deviation IQ score ☐ *T* score ☐ Scaled score

Composite 1: _____

Maximum norm-referenced score: _____ ☐ Ceiling violation

Minimum norm-referenced score: _____ ☐ Floor violation

Internal consistency reliability: _____ ☐ Under .95

Test–retest reliability: _____ ☐ Under .90

Validity evidence: ☐ Content ☐ Response processes ☐ Internal structure
☐ External relations ☐ Consequences

Composite 2: _____

Maximum norm-referenced score: _____ ☐ Ceiling violation

Minimum norm-referenced score: _____ ☐ Floor violation

Internal consistency reliability: _____ ☐ Under .95

Test–retest reliability: _____ ☐ Under .90

Validity evidence: ☐ Content ☐ Response processes ☐ Internal structure
☐ External relations ☐ Consequences

Composite 3: _____

Maximum norm-referenced score: _____ ☐ Ceiling violation

Minimum norm-referenced score: _____ ☐ Floor violation

Internal consistency reliability: _____ ☐ Under .95

Test–retest reliability: _____ ☐ Under .90

Validity evidence: ☐ Content ☐ Response processes ☐ Internal structure
☐ External relations ☐ Consequences

Composite 4: _____

Maximum norm-referenced score: _____ ☐ Ceiling violation

Minimum norm-referenced score: _____ ☐ Floor violation

Internal consistency reliability: _____ ☐ Under .95

Test–retest reliability: _____ ☐ Under .90

Validity evidence: ☐ Content ☐ Response processes ☐ Internal structure
☐ External relations ☐ Consequences

(continued)

Composite 5: _____

Maximum norm-referenced score: _____ ☐ Ceiling violation

Minimum norm-referenced score: _____ ☐ Floor violation

Internal consistency reliability: _____ ☐ Under .95

Test–retest reliability: _____ ☐ Under .90

Validity evidence: ☐ Content ☐ Response processes ☐ Internal structure
☐ External relations ☐ Consequences

Composite 6: _____

Maximum norm-referenced score: _____ ☐ Ceiling violation

Minimum norm-referenced score: _____ ☐ Floor violation

Internal consistency reliability: _____ ☐ Under .95

Test–retest reliability: _____ ☐ Under .90

Validity evidence: ☐ Content ☐ Response processes ☐ Internal structure
☐ External relations ☐ Consequences

Subtest scores

Ceiling violations (list subtests; see tables in Chapter 7 for indications): _____

Floor violations (list subtests; see tables in Chapter 7 for indications): _____

Item gradient violations (list subtests' and the raw scores associated with violations): _____

Notes:

CHAPTER 6

Interpreting Intelligence Test Scores

According to the *Merriam–Webster Dictionary*, the verb *interpret* means "to explain or tell the meaning of" and "to present in understandable terms," and the noun *interpretation* means "the act or the result of explanation." This chapter focuses on interpretation of an examinee's responses and resulting scores. It targets interpreting data from qualitative and quantitative perspectives, using interpretive strategies that are based on an understanding of the nature of and relations between cognitive abilities, and relying on the scientific research base. This chapter begins with discussion of general frameworks applied during interpretation, and it continues with discussion of scores commonly yielded by intelligence tests. It ends with a description of strategies for score interpretation that are consistent with standards for evidence-based practice.

FOUNDATIONS FOR INTERPRETATION

Interpretive frameworks for understanding performance on intelligence tests generally fall along two dimensions (see Floyd & Kranzler, 2012). One dimension reflects the contrast between interpretation of *qualitative* and *quantitative* data. Qualitative approaches draw meaning from differences in the kinds of behaviors exhibited by those being tested. Narrative descriptions of odd test behaviors, recordings of vocalizations, and patterns of errors across items are examples of qualitative data. In contrast, quantitative approaches draw meaning from the numbers derived from assessment. IQs and frequency counts are examples of quantitative data.

The second dimension reflects the contrast between *nomothetic* and *idiographic* interpretations (Allport, 1937). Nomothetic interpretations are based on comparison of attributes of an individual to a larger group (e.g., a *norm group*). These interpretations are interindividual in nature and reflect the relative standing of the individual on some measurement; they are norm-based. Idiographic interpretations are based on understanding the attributes of an individual—all other things being equal. For example, reviewing a child's developmental history and considering the types of errors made during completion of an intelligence test subtest are consistent with the idiographic approach. Furthermore, idiographic interpretations may also reflect the comparison

of some of an individual's attributes to his or her other attributes. These interpretations are intra-individual in nature (a.k.a. person-relative). In this section of the chapter, we consider these two dimensions in concert as we organize discussion of interpretive methods and specific score types.

Qualitative Idiographic Approaches

There is a tradition of applying qualitative and idiographic approaches to interpretation of intelligence tests and related assessment instruments, and this tradition is particularly strong in clinical neuropsychology (Semrud-Clikeman, Wilkinson, & Wellington, 2005). These approaches make use of the extremely rich data that stem from testing sessions. In essence, every purposeful behavior and involuntary response by the person being tested produces qualitative and idiographic data that can be interpreted as meaningful and relevant to the presenting problem. An examinee's trembling voice, shaking hands, and sweat on the brow may indicate excessive anxiety and perhaps fear. Refusal to remain seated during testing, excessive fidgeting, and impulsive responding to test items may indicate attention-deficit/hyperactivity disorder (ADHD). Vocalized self-derision (e.g., "I am so stupid; I just can't do math") and statements indicating hopelessness (e.g., "I'll never be good at reading") may indicate a depressive disorder. Asking for orally presented items to be repeated may indicate a hearing problem, just as an examinee's moving his or her face closer to or farther away from words printed on a page may indicate a vision problem. Rare behaviors by examinees, such as frequent rotations on block construction tasks, cursing, and bizarre statements, may also reflect brain injuries or serious mental disorders. On the other hand, an examinee's behaviors may indicate adaptive strategies. For example, examinees may benefit from counting on their fingers or "writing" a math problem with a finger on the table. Others may benefit from closing their eyes to screen out distractions or using rehearsal strategies when asked to repeat orally presented information.

It is important to consider qualitative idiographic data during completion of a comprehensive assessment. Such data aid in understanding the whole child and in generating ideas for intervention strategies that might otherwise be overlooked if only quantitative data are considered. The one-on-one interactions between examiner and examinee during testing sessions provide opportunities—perhaps unique ones—for carefully studying behaviors that may be meaningful and relevant to the presenting problem. For example, an adolescent's facial tics or petit mal seizures may have previously been overlooked by teachers, parents, and other adults in his or her life, and keen observations of the adolescent during testing may lead to identification of associated disorders. The potential of these data for generating hypotheses to be evaluated via more objective and ecologically valid methods cannot be ignored.

Interactions between examiner and examinee during testing sessions provide opportunities to carefully study behaviors that may be meaningful and relevant to the presenting problem.

In fact, there is little to no evidence supporting the validity or utility of such qualitative idiographic approaches for assessing high-incidence disabilities. In addition, there is a real threat of invalid inferences and other decision-making errors based on the nature of the assessment methods underlying them. These data are typically most useful in a highly limited manner for informing statements about the validity of the quantitative scores yielded by intelligence tests (see Chapter 4).

Quantitative Idiographic Approaches

Quantitative and idiographic methods to interpretation are relatively uncommon in the intelligence-testing literature. *Raw scores* from individual subtests also fall into this category; the adjective *raw* refers to the fact that the score is not changed in any way. The raw scores most typically mentioned in test manuals and research represent the number of correct responses, the number of points earned, or the number of errors made during a subtest; they are sums calculated from item-level scores. Everyone understands when these raw scores are used to calculate *percentage correct* (another quantitative idiographic variable). For example, a child who correctly completes 16 of 20 items receives a percentage correct score of 80%. Most intelligence test subtests include items that increase in difficulty, so it is extremely uncommon to calculate and report percentage correct scores based on item scores. Percentage correct scores should not be confused with percentile rank values, which are discussed later in this chapter.

There are substantial limitations in interpreting raw scores. Because item-based scores tend to be highly unreliable, notable variation in the total number of raw score points due to error should also be expected, and there is no easy way to adjust these raw scores to control for this error. By nature, raw scores also lack meaningfulness in determining whether the performance yielding them can be considered "normal" or "abnormal." For instance, a raw score of 10 on a subtest targeting vocabulary knowledge may be well above average for a 5-year-old in comparison to same-age peers, but well below average for an 18-year-old. The process of norm referencing addresses this limitation; raw scores are the foundation for the vast majority of the other scores in the quantitative nomothetic category.

In recent decades, advances in psychometrics, particularly item response theory (see Embretson & Reise, 2000), have permitted for the transformation of raw scores into more refined scores that have equal units along their scales. Examples include *ability scores* from the Differential Ability Scales—Second Edition (DAS-II; Elliott, 2007); *change-sensitive scores* from the Stanford–Binet Intelligence Scales, Fifth Edition (SB5; Roid, 2003); and *W scores* from the Woodcock–Johnson III Tests of Cognitive Abilities (WJ III COG; Woodcock, McGrew, & Mather, 2001; see Chapter 7 for more details about these tests). These scores are absolute in measuring the targeted ability (unlike norm-referenced scores) and may be useful in examining change in the abilities underlying test performance across time. However, most test users do not attend to and interpret these scores. Although such scores offer promise in evaluating change over time (e.g., in progress monitoring; see Chapter 9), intelligence tests were not developed for this purpose.

As described briefly in Chapter 4 and consistent with Feuerstein's test–teach–test method mentioned in Chapter 13, *testing of limits*, a technique conducted after completion of the standardized administration, can facilitate another type of quantitative idiographic interpretation. For example, if a child seems to misunderstand a lengthy question during standardized administration of a test of knowledge, you could reword that question—clarifying the points of confusion—and determine whether your rewording elicits the response that was targeted. Typically, performance after this testing-of-limits "intervention" is compared informally to that of the original performance—in a qualitative manner. Such methods have been standardized in some tests, such as the Wechsler Intelligence Scale for Children—Fourth Edition Integrated (WISC-IV Integrated; Kaplan et al., 2004), as described in Chapter 4; however, differences between the original and follow-up testing conditions are tainted by confounds, such as carryover effects and guessing.

Finally, this category includes *ipsative analysis*, an intraindividual or person-relative comparison of scores that is sometimes called *profile analysis*. Although the scores included in an ipsative analysis are typically based on reference to a norm group, it is the comparison of these scores for an individual that makes them idiographic. These scores may be subtest scores or composite scores, but the subtest-level ipsative analysis has historically been most common. Ipsative analysis has typically involved calculation of the average across subtest scores and subsequent comparison of each subtest score to that average. When the difference between a subtest score and the average of other scores is substantial and the subtest score in question is higher than the average, the subtest score is interpreted as indicating a *relative strength*. In contrast, when the difference between a subtest score and the average of other scores is substantial and the subtest score in question is lower than average, the subtest score is interpreted as indicating a *relative weakness*. In some schemes, these labels can be applied only if the difference between the average across scores and the score in question is not likely to have happened by chance (i.e., is statistically significant), and if such a difference can be considered rare.

Despite the intuitive appeal of conducting an ipsative analysis of subtest scores, a sizeable body of evidence has indicated that this practice is fraught with error and that it does not contribute to diagnostic or treatment utility (for a review, see Watkins, Glutting, & Youngstrom, 2005). Unlike the overall score on intelligence tests, patterns of relative strengths and relative weaknesses for individuals tend to have rather poor stability. In addition, as mentioned in Chapter 1, they add little to prediction of important criteria beyond psychometric *g*. Moreover, at the current time, the results of ipsative analysis do not improve diagnostic accuracy and do not help with treatment planning. Not only are patterns of subtests ineffective for group differentiation (e.g., those with and without specific learning disability [SLD]), but the average profile within any diagnostic category does not characterize the profiles of every member of that group. Specific profiles of a particular diagnostic group also may be characteristic of members in other groups. Although conducting ipsative analysis at the composite score level has been increasingly recommended over the past decade (e.g., see Flanagan & Kaufman, 2009; Lichtenberger & Kaufman, 2009)—in part due to the higher reliability estimates supporting composite scores than subtest scores, as well as closer ties to prominent models of cognitive abilities—the most salient criticisms of subtest-level profile analysis also apply to composite score profiles. Composite-level ipsative analysis methods have been evaluated far less than those at the subtest level, but at present there is very little evidence to suggest that the subtest or composite score differences on intelligence tests can be used to improve decisions about individuals.

Conducting an ipsative analysis of subtest scores is fraught with error and does not contribute to diagnostic or treatment utility.

Qualitative Nomothetic Approaches

Interpretive methods that focus on the kinds of behaviors—not numerical counts of them—exhibited during testing, and that compare these behaviors to some group-based standard, are extremely uncommon. In fact, by definition, all nomothetic approaches are necessarily quantitative; however, a few interpretive approaches come closest to representing this category. For example, examiners using the WISC-IV Integrated (Kaplan et al., 2004) can (1) observe and record qualitative behaviors, such as requests for item repetition, pointing responses, and use of extra

blocks and breaks in the final configurations; and (2) determine how rare such behaviors are when compared to those observed in the normative group. In general, such interpretive approaches enhance the meaningfulness of more basic qualitative observations and are intriguing—but in the absence of a clear pattern of external correlates, test users should probably avoid them.

Quantitative Nomothetic Approaches

Intelligence test interpretation has strong quantitative and nomothetic foundations. As described in detail in Chapter 2, the foundation for understanding and measuring intelligence in the practice of psychology is the normal curve (see Figure 2.1). The normal curve represents very well the patterns of individual differences across cognitive abilities of varying levels of generality (Carroll, 1993). Individual differences in cognitive abilities are inferred from the deviation of the individual's performance from the average performance determined from testing large groups of individuals with similar characteristics (e.g., age). For intelligence tests, the expectation is that this group—the *norm group*—is large and representative of the population as a whole (see Chapter 5). Below, we address scores derived from such groups.

Age-Level and Grade-Level Referenced Scores

Some, but very few, modern intelligence tests yield both age and grade equivalent scores. Interpretation of these scores is reflective of a quantitative and nomothetic approach, but they differ in many ways from both (1) a focus on individual differences (as described earlier in the chapter) and (2) the most commonly used norm-referenced scores, standardized scores and percentile ranks (described later).

AGE EQUIVALENTS

Age equivalents typically represent the age level at which the typical score (i.e., the mean or median score) on a subtest is the same as an examinee's raw score. More specifically, they reflect the level of development at which the typical raw score earned by a particular age group equals the raw score earned by an individual. Age equivalents are typically expressed in years and months with a hyphen or dash between them, such as 7-1, 8-11, and 10-10. Traditionally (but erroneously), such scores have been interpreted as reflecting "mental age," and inferences indicating that a child has the "mind of a 2-year-old" have been drawn from them.

Age equivalents are gross representations of an individual's level of development inferred from cross-sectional patterns of scores from a norming sample. The comparisons across age levels are statistical and not necessarily related to particular developmental milestones; age equivalents do not reference criteria that should be met at a certain age (e.g., being able to speak in three-word sentences by age 2 or being able to identify all the continents by age 9). Because the reference point at every age level is the mean or median score for that small group, age equivalents also do not reference the range of individual differences in abilities at every age level. In the same vein, they do not indicate that about half of children at every age level (e.g., at 7 years, 1 month) would obtain a higher raw score and that about half of children at that age level would obtain a lower raw score. Nor do they indicate how deviant a child is from expectations based on comparisons to

same-age peers. Finally, like the raw scores on which they are based, age equivalents have unequal units along their scale, and are prone to increases or decreases due to error in measurement. For example, a 1-point increase in a raw score may mean a 2-year increase in age equivalents at some age levels and 1/3-year increase at other age levels. Although well-scaled items (due to improved scaling technology) have lessened problems with unequal units along age equivalent scales, these scores are particularly prone to overinterpretation because of the sensitivity of age equivalents to error in measurement.

GRADE EQUIVALENTS

Like age equivalents, grade equivalents typically represent the grade level at which the typical score on a subtest is the same as an examinee's raw score. More specifically, they reflect the particular point in the school year at which the typical raw score earned by a grade-based group equals the raw score earned by an individual. Grade equivalents are typically expressed in grade levels and months of the school year (10 months at most are considered; 0 is September), with a decimal point in between them, such as 7.1, 8.9, and 10.5. (If the value on the right side of the point exceeds 9, the score is probably an age equivalent.) Grade equivalents are more frequently interpreted when yielded by achievement tests than when yielded by intelligence tests. For example, one might conclude (albeit erroneously) that "My first grader is reading on a fifth-grade level."

Like age equivalents, the comparisons across grade levels and months of the school year are statistical and not content-related; grade equivalents do not reference criteria that should be met in a certain grade or school curriculum. They also do not indicate that about half of children at the grade level associated with the score in question would obtain a higher raw score and that about half of children at that grade level would obtain a lower raw score. Nor do they indicate how deviant a child is from expectations based on comparison to same-grade peers. They also have unequal units along their scale. Furthermore, their use assumes that growth in abilities is consistent across the academic year and across school years, and this is clearly not the case in the achievement areas (Nese et al., 2012; Skibbe, Grimm, Bowles, & Morrison, 2012). Grade equivalents are not produced by most intelligence tests; of all the intelligence tests described in Chapter 7, only the WJ III COG (Woodcock et al., 2001) produces grade equivalents.

Scores Derived from Age- and Grade-Based Norms

The most commonly used and useful scores from intelligence tests represent individual differences in abilities inferred from the deviation of the individual's performance from the average performance of a group to which they belong. They are *standardized scores* and *percentile ranks*, and they overcome many of the limitations of age equivalents and grade equivalents.

NORM GROUPS

As described in Chapter 5, norm groups are the products of field testing, standardization, and norming, and they allow for a better understanding of typical levels of performance and variation around it at different levels of development. Age-based norms are the most common type of norms used with intelligence tests. Some intelligence tests cover an extremely broad age range (from toddlerhood to late adulthood), whereas others cover a more narrow age range. In traditional norming

procedures, as discussed in Chapter 5, a sizeable sample of volunteers at every age level completes the intelligence tests under standardized conditions. From their performance, descriptive statistics (e.g., means and standard deviations) are obtained; each raw score is transformed into a standardized score (sometimes through an intermediate score based on item response theory); and norm tables are constructed to link those raw scores to the norm-referenced scores for test users. First, we discuss the general class of norm-referenced scores called *standardized scores*, and then we discuss the associated *percentile ranks*.

STANDARDIZED SCORES

Standardized scores stem from the comparison of raw scores (or sums of any type of scores) to a mean score from a group with reference to the standard deviation of that group. Most graduate students and professionals with undergraduate training in statistics are familiar with the formula for calculating z scores as part of their training in use of statistical tests. For any distribution, z scores have a mean of 0 and a standard deviation of 1, and they typically range from –3.0 to 3.0. Positive z scores indicate that the raw score in question is higher in magnitude than the mean for the sample, and negative z scores indicate that the raw score in question is lower in magnitude than the mean for the sample.

Standardized scores stem from the comparison of raw scores (or sums of any type of scores) to a mean score from a group with reference to the standard deviation of that group.

A z score for an individual can be calculated as a difference between an obtained score *(X)* and a sample mean *(M)* divided by the standard deviation *(SD)* from the sample.

$$z = \frac{X - M}{SD}$$

For example, if the average raw score across 100 third graders who completed a vocabulary subtest was 20 and the standard deviation for this group was 4, the following z scores would be produced.

- A child who receives a score of 20 receives a z score of 0.

$$z = \frac{X - M}{SD} = \frac{20 - 20}{4} = \frac{0}{4} = 0$$

- A child with a raw score of 28 receives a z score of 2.0.

$$z = \frac{X - M}{SD} = \frac{28 - 20}{4} = \frac{8}{4} = 2$$

- A child with a raw score of 14 receives a z score of –1.5.

$$z = \frac{X - M}{SD} = \frac{14 - 20}{4} = \frac{-6}{4} = -1.5$$

All standardized scores are calculated in essentially the same manner, but the means and standard deviations of these scores (not of the original raw score distributions) are somewhat arbitrarily set. As a result of such standardization of raw scores, each standardized score represents its proximity of the mean of the referenced norm group. As a result, individual differences in the targeted ability are represented in these standardized scores. The process of standardizing the score (making the raw scores higher and lower relative to their own mean) allows for comparisons to be made between scores from the same intelligence test or across scores from different tests. In contrast to age equivalents and grade equivalents, standardized scores have more equal units along the scale.

The most common type of standardized scores yielded by intelligence tests are *deviation IQ scores*; they are often called *standard scores*. They have a mean of 100 and a standard deviation of 15. Most IQs and closely associated composite scores across intelligence tests are scaled by using these deviation IQ scores (see Chapter 7). Intelligence tests are typically limited to a range of deviation IQ scores from four standard deviations below the mean to four standard deviations above the mean (i.e., 40–160). Score range labels often indicate that deviation IQ scores from approximately 10 points above and below 100 are in the *Average* range, and additional score range labels are applied to scores approximately 10 points above and below the Average range (e.g., 81–90 = *Low Average* and 110–119 = *High Average*). More information on these score labels is provided in Chapter 8.

In order to differentiate subtest scores from IQ and other composite scores, intelligence test authors have often had them yield other types of standardized scores. Some intelligence test subtests yield *T scores*, which have a mean of 50 and a standard deviation of 10; others employ *scaled scores*, which have a mean of 10 and a standard deviation of 3. As is evident in Figure 2.1 (in Chapter 2), these standardized scores are generally interchangeable with the deviation IQ scores. We believe that students should be able to quickly recall the means for deviation IQ scores, *T* scores, and scaled scores, as well as scores associated with one standard deviation below the mean (i.e., 85, 40, and 7, respectively), two standard deviations below the mean (i.e., 70, 30, and 4, respectively), one standard deviation above the mean (i.e., 115, 60, and 13, respectively), and two standard deviations above the mean (i.e., 130, 70, and 16, respectively).

Although these three types of standardized scores are typically interchangeable, the ranges of these scores and their scaling do make a difference. For example, as evident in Figure 2.1, scaled scores are limited to a lower end range of 3⅓ standard deviations (i.e., a scaled score of 1), whereas deviation IQ scores and *T* scores can go many standard deviations lower before reaching a score of 0, the end point on the scale. In addition, scores with smaller standard deviations (e.g., scaled scores) rather than larger standard deviations (e.g., deviation IQ scores) lead to more abrupt score "jumps" between one area of the normal curve and another area with one score point difference. Because standardized scores tend to be limited to whole numbers, each scaled score spans ⅓ of a standard deviation, each *T* score 1/10 of a standard deviation, and each deviation IQ score 1/15 of a standard deviation. That is, there are far wider gaps in the normal curve between scores for scaled scores than for deviation IQ scores. For example, as evident in Figure 2.1, the amount of area under the curve between a scaled score of 10 and a scaled score of 13 is almost 35%; in contrast, the difference between a deviation IQ score of 100 and a deviation IQ score of 103 is only about 3%. We consider these limitations in subtest scaling (vs. IQ and composite scores that tend to use deviation IQ scores) when recommending that subtest scores not typically be interpreted.

PERCENTILE RANKS

Percentile ranks also represent the degree of deviation of a score from the mean of the referenced group. Percentile ranks provide a representation of an examinee's relative position (following rank order) within the norm group. In particular, they indicate the percentage of the norm group that scored the same as or lower than the examinee. Percentile ranks are expressed as percentages (with a range from 0.1 to 99.9) that reflect the area under the normal curve (see Figure 2.1). Like standardized scores, they allow comparisons to be made between scores within and across intelligence tests. We believe that students should be able to quickly recall the percentile ranks at the upper and lower end of the Average range (i.e., percentile ranks of roughly 25 and 75) and at two standard deviations above and below the mean (i.e., percentile ranks of roughly 2 and 98). In addition, they should know the standardized scores roughly associated with certain percentile ranks, such as the 10th percentile (81, 37, and 6 for deviation IQ scores, *T* scores, and scaled scores, respectively) and the 90th percentile (119, 63, and 14 for deviation IQ scores, *T* scores, and scaled scores, respectively).

Percentile ranks provide a representation of an examinee's relative position (following rank order) within a group.

Percentile ranks share some limitations with standardized scores, and they have some limitations of their own. Like some standardized scores, their range is somewhat limited. At least when percentile ranks as whole numbers are considered, they reach a ceiling at the 99th percentile and a floor at the 1st percentile. Scores higher or lower than these points will yield percentile rank differences of a fraction of a point (e.g., 99.53, 99.62, 99.69, 99.74, etc.) and at another point (approximately $3^2/_3$ standard deviations above or below the mean), the percentile ranks must be differentiated by thousandths of a point. These decimal fractions are challenging to explain clearly, but we discuss them further in Chapter 8. In contrast to standardized scores, percentile ranks have unequal units along their scale. As evident in Figure 2.1, a 10-point difference between percentile ranks near the mean (e.g., between percentile ranks of 40 and 50) would indicate small differences in standardized scores (i.e., deviation IQ scores of about 96 and 100) and would thus be of little importance, whereas in the tails of the normal curve, a 10-point difference between percentile ranks would indicate substantial and perhaps important differences.

CONFIDENCE INTERVALS FOR STANDARDIZED SCORES AND PERCENTILE RANKS

As discussed in Chapter 5, reliability in measurement is an essential feature to consider in selecting tests and determining which test scores are likely to yield the best information during assessment. At the score level, reliability is akin to accuracy in measurement, and the flip side of reliability is *measurement error*. Measurement error is represented as the difference between a reliability estimate (especially an internal consistency reliability coefficient) and unity (i.e., 1.00). Such error can be incorporated into interpretation by considering *confidence intervals* surrounding scores.

Just as we commonly hear the results of polls followed by mention of the margin for error (e.g., "plus or minus 5 percentage points"), many norm-referenced scores can be surrounded by bands of error. The width of the bands of error tell us how much to expect a child's obtained score (the single norm-referenced score yielded by the test) to vary from his or her "true score" if the child

was administered the same test repeatedly (assuming no practice effects, fatigue effects, or the like). Following a prominent analogy, picture two archers of varying skill levels shooting at a target on a calm day. Both archers aim at the target's center, the bull's-eye, and both shoot 10 arrows. Because neither archer is perfect and because environmental influences affect the trajectory of arrows, we would expect the arrows to be scattered about the target in a predictable pattern—with a higher density in the middle of the target and a lower density toward the borders of the target. However, these patterns will be likely to differ for the two archers. The more skilled archer will hit the bull's-eye occasionally and will have a greater number of arrows near it and relatively few arrows near the border. In contrast, the less skilled archer will hit near the bull's-eye with a couple of arrows but will scatter others around the target, with a greater number of arrows near the border and perhaps some arrows missing the target altogether.

The archers' skill level is equivalent to reliability. Less skilled archers scatter their arrows across a wider range of segments of the target than more skilled archers, and less reliable tests produce scores that are more likely to vary across the range of hypothetical administrations of the test than the scores of more reliable tests are. In addition, to continue with this metaphor, it is possible (although highly improbable) that the more skilled archer will miss the target altogether with one arrow, while hitting the bull's-eye with another and coming very close to it with the vast majority of their other arrows. In such a case, we could hypothesize that the one errant shot is due to "bad luck" or some extraneous, distracting influence, but we would understand that although it is not likely to happen again, such a bad shot is not outside the hypothetical range of possible shots. However, the shot does not appear to be representative of the archer's true skill level. Similarly, the less skilled archer may hit the bull's-eye while missing the target altogether across all other shots; this perfect shot could be considered a random, improbable occurrence (i.e., really good luck) within the hypothetical range of possible shots and not representative of the archer's true skill level. The most accurate estimates of an archer's skill level would stem from shooting multiple arrows across multiple trials. In the same vein, the most reliable scores stem from multiple items from multiple subtests considered together in a composite score.

In order to better understand confidence intervals, you must also understand the standard error of measurement (SE_m). The SE_m is based in part on the reliability coefficient (most commonly, the internal consistency reliability coefficient) described in Chapter 5, and the other key variable is the standard deviation of the score in question. Confidence intervals are most commonly reported for standardized scores (and specifically for deviation IQ scores), so we focus on them here. The formula for the SE_m requires that an estimate of error be obtained by subtracting the reliability coefficient from 1.0. The square root of this value is obtained, and it is multiplied by the standard deviation of the score (in the case of a deviation IQ score, 15).

If the reliability of a score is .97, and its standard deviation is 15, the formula can be applied in this manner:

$$SE_m = 15\ (\text{SQRT}(1 - .97))$$
$$SE_m = 15\ (\text{SQRT}(.03))$$
$$SE_m = 15\ (.1732)$$
$$SE_m = 2.60$$

This SE_m is the standard deviation of the distribution of hypothetical *true scores* around, most frequently, the obtained score. True scores, according to classic test theory, are hypothetical enti-

ties that represent—with perfect accuracy—the targeted construct. If we consider the same normal curve discussed earlier in this chapter and presented in Figure 2.1, the same "rules" for area under the normal curve apply to the distribution of error around the obtained score. For example, because the SE_m is the standard deviation of the distribution of hypothetical true scores, we know that about 68% of the true scores would fall between 1 SE_m above the obtained score and 1 SE_m below the obtained score. For instance, considering a child's obtained IQ of 100 and a SE_m of 2.6, we can anticipate that the child's true IQ would fall within the range of 97.4 and 102.6 about two out of three times (i.e., 68% of the time) if he or she were able to take the test repeatedly (assuming no carryover effects, etc.).

Knowing the area under the curve, we can use the SE_m values to calculate *confidence intervals*, which represent the range of true scores around the obtained score beyond 1 SE_m. A confidence interval is calculated by multiplying the SE_m by standard deviation units expressed as *z* scores (see our discussion of *z* scores earlier in the chapter). For example, we would multiply the SE_m by 1.65 to obtain the 90% confidence interval, by 1.96 to obtain the 95% confidence interval, and by 2.58 to obtain the 99% confidence interval. For reference, the SE_m of 2.6 produces these confidence intervals:

90% confidence interval: ±4.29
95% confidence interval: ±5.10
99% confidence interval: ±6.71

Our review of intelligence tests in Chapter 7 indicates that most test authors center confidence intervals around an *estimated true score* versus the actual score earned by the examinee (a.k.a. the obtained score). This practice is apparent when the confidence interval values above and below the obtained score are not symmetrical. Estimated true scores (also called *regressed true scores*) are practically always closer to the mean of the population distribution than obtained scores. Obtained scores are more deviant from the mean in part because of error (i.e., chance) deviations due to guessing, "bad luck," or the like. Estimated true scores are derived by (1) multiplying the difference between the obtained score and the mean by the same reliability coefficient used to calculate the SE_m, and (2) adding this product to the mean. Because no measurement has perfect reliability, the difference from the population score mean (e.g., 100) for the estimated true score is always smaller in magnitude than the difference from the population score mean of the obtained score; it is a product of multiplying this difference by a value less than 1 (e.g., .95). In addition, as evident in Chapter 7, almost every prominent intelligence test produces confidence intervals based on adjustments to the SE_m by multiplying it by the same reliability coefficient to produce the *standard error of the estimated true score* (SE_E). The SE_E values are always smaller than the SE_m values, for the same reason explicated for the estimated true score: We are multiplying it by a number that is less than 1.0.

In many ways, confidence intervals seem to produce a paradoxical effect. In using them, we seem to sacrifice precision (as represented by a single obtained score) for confidence in estimating the true score. The more confident we are that the true score falls within the SE_m or SE_E interval (e.g., 95% confidence), the less precise we are; conversely, the less confident we are (e.g., 68% confidence), the

Confidence intervals seem to produce a paradoxical effect. In using them, we seem to sacrifice precision for confidence in estimating the true score.

more precise we are. Confidence intervals show us that we may, in fact, better represent the abilities we have targeted by expressing performance as a range of scores. We encourage you to employ confidence intervals when reporting results (see Chapter 8).

ADVANCED QUANTITATIVE INTERPRETIVE METHODS

So far in this chapter, we have addressed general interpretive frameworks and the most basic score-based interpretations. This section is devoted to the most prominent methods for understanding patterns of quantitative nomothetic data yielded by tests. With the increasing sophistication of intelligence tests—which now yield a greater number of subtest and composite scores, and which target abilities at various levels of the three-stratum theory (Carroll, 1993)—test users may be overwhelmed with the wealth of information yielded by a single test. In order to make sense of all this information, we offer some general rules of thumb to follow and our KISS model of interpretation.

The First Four Waves of Interpretation

One of our reasons for writing this book is our belief that measurement and understanding of cognitive abilities in applied psychology have never been stronger. This greater understanding is reflected in the progression of prominent models of test score interpretation, detailed in Kamphaus, Winsor, Rowe, and Kim's (2012) description of the four waves of test interpretation. The first wave reflected the quantification of the general level of intelligence. During this wave of interpretation, a focus on psychometric *g* as measured by IQs was predominant, and individuals were placed in rank order primarily on a single scale measuring this ability. With advancing technology in the form of intelligence test subtests that each yielded norm-referenced scores, the second wave of interpretation emerged. It was characterized by clinical analysis of patterns of subtest scores in addition to the IQ, and idiographic and qualitative interpretations were highlighted. More specifically, subjective and idiosyncratic analysis of subtest profiles guided by psychoanalytic theory and clinical lore typified this approach. The third wave highlighted advances in psychometrics that guided (1) the application of factor analysis to test interpretation, and (2) ipsative analysis of subtest scores within the profiles of individuals using statistical tests to determine significant differences. In particular, this wave was reflected in greater consideration of the shared measurement properties of subtests (used to form specific-ability composites) and their unique components, as well as the evidence supporting the interpretation of profiles of IQ, composite, and subtest scores.

We believe that measurement and understanding of cognitive abilities in applied psychology have never been stronger.

Finally, the fourth wave reflects the application of research-based models of cognitive abilities to the development of tests and score interpretation. The most prominent models guiding the fourth wave have stemmed from the theoretical models of Carroll (1993) and Horn (1991). Despite their differences, these two models have been integrated in the Cattell–Horn–Carroll (CHC) theory (as described in Chapter 1), and it has been promoted as the most well-supported and sophisticated psychometric model of intelligence (Newton & McGrew, 2010; Schneider &

McGrew, 2012). Furthermore, authors of prominent intelligence tests (e.g., Kaufman & Kaufman, 2004a; Woodcock et al., 2001) have developed tests and produced scores representing a number of the broad abilities described in the CHC theory.

We agree with Kamphaus et al. (2012) and others that this fourth wave, and the convergence of research and professional opinion on models like the CHC theory, afford the test user at least three benefits. First, scores from all intelligence tests can be interpreted by using these models; test users need not rely primarily on test-specific models, which may vary substantially from one test to another. Second, test users can select an intelligence test that best meets their needs. That is, their test selection need not be based only on their theoretical leanings or the differences in the quality of the tests per se. Instead, they can choose tests according to the age level and limitations of their typical clients, the clients' referral concerns, the breadth of specific-ability coverage, the time of administration, and other individual preferences. Third, they can benefit from training in understanding a general theoretical model and from books that promote its application to test score interpretation.

Making Meaning in the Fourth Wave and Considerations for a Fifth Wave

The most common approach to interpreting scores yielded by an intelligence test has been the *successive-levels approach*. In one classic model, Sattler (2008) has suggested that six levels of interpretation be addressed. First, the IQ is interpreted. Second, scores from lower-order composites are interpreted. Third, scores from subtests contributing to each lower-order composite are interpreted in isolation and in comparison to the average across subtests contributing to their respective lower-order composite. Fourth, scores from subtests are compared across the entirety of the test (including in meaningful pairs). Fifth, patterns of item-level scores are evaluated for each subtest. Finally, a qualitative idiographic analysis of item-level responses and other behaviors during the test sessions is conducted. In order to maximally apply the fourth wave of interpretation and to make the transition to more evidence-based practices, we encourage test users to attend to two psychometric considerations—reliability and *generality*–and we offer some additional points that may be considered in a fifth wave of interpretation.

Reliability

Following the principles discussed in Chapter 5, we know that scores closest to the item level (item-level scores and subtest scores) should not attract our attention to the extent that scores stemming from aggregation across numerous items and several subtests should. Composite scores that exceed minimal standards for internal consistency and test–retest reliability should be the focus of our attention. As apparent for every intelligence test evaluated in Chapter 7, stratum III composites (i.e., IQs) yield the highest reliability coefficients, followed by stratum II composites (e.g., factor indexes). Because reliability constrains validity, the most reliable scores from intelligence tests will yield the most dividends for predicting educational, occupational, and other life outcomes and addressing the most pressing referral concerns. Be choosy about which scores you interpret, and feel comfortable ignoring scores if they do not meet the highest standards of reliability and validity. Just because the test or its scoring software produces a score, you need not interpret it.

Generality

In a manner similar to reliability, we know that when scores from numerous items and multiple subtests are aggregated, the influences of specific item content and specific mental processes largely cancel out one another. As a result, more general abilities are measured. In contrast, based on our perceptions and common sense, intelligence test subtests measure abilities in very specific ways. For example, a subtest requiring examinees to respond to orally presented words by orally providing definitions of these words seems to measure vocabulary knowledge versus a more general ability associated with thinking abstractly and problem solving. Other subtests may also appear to measure the vocabulary knowledge via different methods, such as having the examinee (1) hear an orally presented word and point to a picture that represents the word, or (2) view pictures and name what they depict. These related subtests may share a common element—regardless of the assessment method used—that produces higher or lower scores across individuals. For example, that common element may, in fact, be vocabulary knowledge, which is a stratum I (or narrow) ability (Lexical Knowledge). These subtest scores also tend to be highly correlated with subtests measuring knowledge of cultural phenomena and the ability to reason using words. The vocabulary knowledge subtests and other related subtests measuring breadth of knowledge seem to tap into a more general ability that could be called Crystallized Intelligence at stratum II. In this case, the individual subtest characteristics—and even vocabulary knowledge more generally—reflect only pieces of the larger puzzle, because they are more specific than this stratum II (or broad) ability that produces higher and lower scores across similar subtests.

As described in Chapter 1, it is a known fact that almost every measure of cognitive abilities tends to be positively related to most every other measure of cognitive abilities, which suggests that they are measuring the same thing. How can this be? We can see how vocabulary knowledge subtests measure the same thing and how, on a broader level, Crystallized Intelligence subtests measure the same thing, but how is it possible that vocabulary knowledge subtests measure the same thing as subtests requiring the construction of designs with blocks or the identification of abstract patterns across images on a page? It's *uncommon sense*: Believing the evidence produced through more than 100 years of scientific research (without necessarily experiencing it; see Lilienfeld, Ammirati, & David, 2012) allows us to conclude that there is an extremely general superfactor, the *g* factor, that accounts for score variation across all such subtests. Subtests that look vastly different are apparently measuring the same things to a surprising degree.

As discussed in Chapter 1, we can conclude that the ability at the highest level of generality is the most powerful influence in producing higher or lower scores on intelligence tests, and that the measures best representing this ability tend to produce the strongest relations with important societal outcomes. As we focus on more and more specific abilities rather than this general ability, the explanatory and predictive power afforded by these specific abilities tends to decrease dramatically. Considerations of both reliability and generality lead us to conclude that the most accurate and most general scores from intelligence tests, the stratum III composites (the IQs), should be the focus of interpretation. Scores with lower levels of reliability and less generality are likely to produce the results that seem to have deep meaning, but research examining errors in clinical decision making

The most accurate and most general scores from intelligence tests, the stratum III composites (the IQs), should be the focus of interpretation.

(in general) and in subtest score interpretation (more specifically) indicates that results from such scores most likely reflect only illusions of meaning (Lilienfeld et al., 2012; Watkins, 2009).

But can't the specific abilities represented in Carroll's three-stratum theory and CHC theory be measured? The answer is yes, but their effects cannot be isolated with accuracy when interpreting intelligence test scores (Gustafsson, 2002; Oh, Glutting, Watkins, Youngstrom, & McDermott, 2004). When we refer to *specific abilities*, we are considering influences on test scores that are independent of the *g* factor (see Chapter 1). In fact, when we are targeting specific abilities per se, variability due to the general factor can be considered construct-irrelevant. (The same may be true for measures of the *g* factor; we certainly do not want lots of specific-ability variance entering into the IQs.) In general, specific-ability variance estimates tend to be relatively low when compared to total variance and general factor variance; those measures that tend to be strongly influenced by the *g* factor tend to have weaker specific-ability estimates. Basically, very few subtests and only some composite scores measure a sufficient amount of specific-ability variance to warrant their interpretation as indicators of those more specific abilities.

We see only three options for interpreting scores as representing specific abilities, and all of these options have serious limitations:

1. *Interpret sufficiently reliable composite scores stemming from two or more subtest scores as representing specific abilities*. This practice is the most common, but it fails to consider that general factor variance in these composite scores is construct-irrelevant when the composite targets specific abilities. Moreover, many highly reliable composites tend to have more variance associated with the general factor than with specific abilities. This method is a very crude way to examine individual differences in specific cognitive abilities; it works primarily if the general factor is ignored during interpretation of these composites.

2. *Interpret scores from sufficiently reliable composites (stemming from two or more subtest scores) in a profile of related scores using an ipsative analysis. When composite scores differ significantly from the profile mean, it indicates that strengths or weaknesses in specific cognitive abilities are at play. If they are, place more emphasis on them during interpretation*. This practice is far less common than the first and it has some potential for carefully formulated profile analysis. However, it is limited in several ways. Because such composites tend to vary in the influence of the general factor on them and in their reliability, the amount of variance associated with specific abilities in each composite also varies substantially. One composite score may appear to be significantly different from the others, indicating a specific-ability strength or weakness, but this may be due in large part to its weaker reliability than the others. Furthermore, (1) these composite strengths and weaknesses are likely to be unstable across time (as that seen with subtests), and (2) we have no scientific evidence of benefits garnered by using this interpretive practice.

3. *Create specific ability factor scores (removing the construct-irrelevant influences of the g factor on them), and interpret them after they are norm-referenced*. We support the creation of such composite scores that represent separate abilities from varying strata of cognitive abilities (e.g., stratum III, stratum II, etc.) and are excited about recent advances in actualizing this promise (Schneider, 2012). At present, however, no intelligence tests produce composites representing specific abilities in this manner. We also suspect that the reliability of these scores will tend to be inadequate because they stem from such small slices of test score variance, but confidence interval

values could be applied to them (Schneider, 2013). At present, we have no good evidence supporting interpretation of such factor scores representing specific abilities and no evidence at all of the benefits stemming from their interpretation. We remain open-minded about the potential of this method and await the development of a strong body of validity evidence before recommending that it be applied during high-stakes testing, however.

A Fifth Wave

We believe that a fifth wave of interpretation will build on the strengths of the fourth wave and that it will rely on evidence-based and empirically supported interpretations that go beyond primarily factor-analytic research. Interpretation in this wave will promote more selective and focused intelligence testing. Consistent with the standards for validity evidence cited in Chapter 5, this wave may require test users to apply only those interpretations that are based on highly focused scientific evidence. This wave will probably lead to a vastly restrictive successive-levels approach that includes, at most, two levels of interpretation: stratum III scores (a.k.a. IQs) and selective lower-order (e.g., stratum II) composite scores that have been shown to have high-quality scientific research supporting their validity (e.g., incremental validity in prediction or treatment utility). Lower levels of interpretation will be cast aside in favor of (1) instruments designed to measures some of the constructs targeted in high-inference interpretation of subtest profiles and item-level responses, and (2) interpretation of only scores supported by a large body of reliability and validity evidence. Item-level scores and scores from intelligence test subtests will not be interpreted. Consistent with the sea change that has led to disregard of idiographic and qualitative interpretations of projective tests, the fifth wave may bring about the end of the wide-ranging qualitative idiographic "clinical interpretations" of intelligence test scores that have lingered on since their heyday in the second wave.

A fifth wave of interpretation will rely on evidence-based and empirically supported interpretations that go beyond primarily factor-analytic research to promote more selective and focused intelligence testing.

The KISS Model in a Fifth Wave

We offer a KISS model to guide interpretation in a fifth wave of interpretation. The acronym KISS may stand for "Keep it simple and scientific" or "Keep it simple, scholar." (Take your pick!) We think that it well represents empirically supported best practice in the field at the current time. We address each component of the model in turn.

We offer a KISS—"Keep it simple and scientific" or "Keep it simple, scholar"—model to guide interpretation in a fifth wave of interpretation.

Interpretation of the Stratum III Composite

Interpretation should begin with consideration of the most reliable, general, and empirically supported score yielded by the intelligence test: the IQ. The preponderance of evidence suggests that

the fifth wave of interpretation should include the same focus as the first wave—consideration of the psychometric *g*—during interpretation. You are typically standing on solid ground when considering an individual's normative level of performance on IQs.

Ipsative analysis of composite and subtests scores has also been conducted to determine the "validity" or "interpretability" of more global scores (e.g., IQs and composites) when there is "scatter" in their constituent parts. When significant scatter is found, the practice has long been to discount the IQ and focus instead on interpreting the individual's profile of subtest or composite scores (Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). In fact, Hale and Fiorello (2001) argued that one should "never interpret the global IQ score if there is significant scatter or score variability" (p. 132); the rationale is that you are figuratively mixing apples and oranges when combining scores that measure different abilities. Recent research, however, has failed to support this contention that the IQ is invalidated when there is significant variability among composite or subtest scores (Daniel, 2007; Freberg, Vandiver, Watkins, & Canivez, 2008; Kotz, Watkins, & McDermott, 2008; Watkins, Glutting, & Lei, 2007). Thus, research supports the interpretation of the IQ even when the scores contributing to it are not consistent.

Interpretation of Stratum II Composites as Measures of Specific Abilities

Interpretation of composite scores as measures of broad cognitive abilities has become increasingly popular in recent years (e.g., Flanagan & Kaufman, 2009; Prifitera, Saklofske, & Weiss, 2005), due in part to the lack of evidence supporting the interpretation of subtests and other problems with subtest-level interpretation (e.g., McDermott & Glutting, 1997; Watkins & Canivez, 2004). We know, though, that we are on more shaky ground now, due to the lessened reliability and lessened generality of these composites. Nevertheless, applying nomothetic quantitative interpretive strategies to these scores seems to paint a more descriptive picture of test performance than focusing on only the IQ. We believe that we should probably engage in selective and cautious interpretation of these composites after considering (1) their reliability (expecting internal consistency values of .90 or higher), (2) the influence of the psychometric *g* and specific abilities on them, and (3) evidence supporting the validity of their interpretation in addition to the IQ. However, without a more sophisticated framework to guide interpretation of these composite scores, we struggle to act on this belief.

Interpretation of Subtests

Most intelligence test subtests have reliability and validity far inferior to those of IQs and stratum II composites; this pattern has been well documented. Furthermore, subtests have two additional weaknesses. First, as mentioned previously, most subtests are scaled (with scaled scores or *T* scores), so that the measurement of ability is less precise than deviation IQ scores. Second, despite their inferior reliability, confidence interval values are extremely rarely applied to subtests; such values would allow for some control over the influences of error on their scores. For all of these reasons, subtest scores and profiles of these scores should not be the focus of interpretation. Norm-referenced subtest scores should be interpreted as indicators of broader and more general abilities—as contributors to IQs and stratum II composites—and not as representations of specific skills or narrow abilities per se.

Interpretation of Item-Level Responses

As we have conveyed elsewhere (Floyd & Kranzler, 2012), qualitative idiographic approaches to interpretation of item-level responses are not supported by enough evidence to recommend their routine use for high-stakes decisions. Moreover, item-level interpretation is fraught by substantial error in measurement. We are prone to critical thinking errors when evaluating item-level responses in isolation and in forming patterns across them. From a practical perspective, there is presently little convincing evidence that the devotion of additional time and effort to consider patterns of errors on intelligence test subtests yields dividends for clinicians or their clients.

SUMMARY

We understand the motives of those trying to obtain the most information possible from intelligence tests; these tests yield a wealth of information. We recognize that qualitative information from intelligence tests can seem to accurately reflect the patterns of strengths and weaknesses displayed by children and adolescents in their everyday lives and that such information can be used as part of a comprehensive assessment to inform effective interventions. Furthermore, we know that specific cognitive abilities have been evidenced by a very large body of research and that they are appropriately included in prominent theories of cognitive abilities. We understand the rationale for reviewing subtest task requirements and inferring what cognitive processes and abilities they measure; we see value in labeling subtests according to these inferences; and we feel that composite scores (and on rare occasions subtest scores) may be useful to interpret. All of this information, however, cannot be treated equally during interpretation, and we must evaluate it with a discerning eye. Our KISS model is a guide to doing so.

Results of recent research do not support the successive-levels-of-analysis approach for the interpretation of intelligence tests. Rather, the interpretation of intelligence tests should focus primarily on stratum III composites (i.e., IQs). Not only are these scores the best estimates of psychometric *g*, but they also are the most reliable and predictive scores yielded by intelligence tests. We encourage test users to focus their efforts on interpreting the norm-referenced scores for IQs and to reference broad confidence intervals surrounding them. We also encourage test users to consider the purpose of their assessment and to tailor their testing toward that purpose, as well as to refer to the evidence base that illuminates empirically supported practices. Interpretation may focus on select stratum II composites if there is strong empirical support for their interpretation as measures of specific abilities, but we do not advocate interpretation of all specific-ability composites yielded by intelligence tests or of specific-ability composite profiles, due to lack of evidence supporting these methods.