The DEATH PENALTY and INTELLECTUAL DISABILITY

Edward A. Polloway, Editor

aaidd
American Association
on Intellectual and
Developmental Disabilities

In the 2002 landmark decision *Atkins v. Virginia,* the U.S. Supreme Court ruled that executing someone with intellectual disability (ID) is a violation of the Eighth Amendment of the U.S. Constitution, which prohibits cruel and unusual punishment. In its 2014 decision in *Hall v. Florida,* the Court ruled that, while states have the right to establish their own rules for handling *Atkins* cases, they are not free to ignore scientific and medical consensus regarding intelligence and the nature and diagnosis of ID. The Court rejected the use of an IQ test score of 70 as a bright-line cutoff for determining ID and ruled that all evidence pertinent to the claim, including adaptive behavior assessments, should be considered.

This AAIDD publication is the authoritative resource on the science that is the basis for the definition of intellectual disability and on the critical issues involved in its diagnosis. The book is a comprehensive and cogent analysis of what is involved in the determination of ID, particularly as it relates to defendants in death penalty cases. AAIDD assembled a group of notable scholars and clinicians to bring the best science to this discussion, and the work of this group is represented in the important resource published here.

aaidd
Knowledge | Support | Empowerment

American Association
on Intellectual and
Developmental Disabilities

95
ISBN 978-1-937604-13-4
5[illegible]995>
9 781937 604134

Published by
American Association on Intellectual and Developmental Disabilities
501 3rd Street, NW, Suite 200
Washington, D.C. 20001
www.aaidd.org

To order
AAIDD Order Fulfillment
501 3rd Street, NW, Suite 200
Washington, D.C. 20001
phone: 202-387-1968 x 216
email: books@aaidd.org
online: http://aaidd.org/publications/bookstore-home

Product No. 4134

Printed in the United States of America

7 Intellectual Functioning

Kevin S. McGrew

This chapter focuses on intellectual functioning, the first prong of the definition of intellectual disability (ID). The initial focus is on a review of the literature on intelligence and the assessment of intellectual functioning. Attention is given to psychometric theories of intelligence, contemporary research, an overview of intelligence test batteries, related measurement concepts, relationships between intelligence test scores, and the relationship between measures of intelligence and adaptive behavior scores. The chapter then provides implications for the implementation of this research in practice.

Summary of Related Research

Psychometric Theories of Intelligence

The *psychometric* approach is the most well-established approach to studying intelligence, dating to Galton's attempt in the late 1800s to measure intelligence with psychophysical measures (Sternberg & Kaufman, 1998). Psychometric theories of intelligence are "based on or tested by scores on conventional tests of intelligence . . . these theories are often, but not always, based on FACTOR ANALYSIS, that is, they specify a set of factors alleged to underlie human intelligence" (capitalization in the original; VandenBos, 2007, p. 754). Psychometric theories have been the most influential approach to measuring and studying intelligence, have generated the most systematic research and, more importantly, have facilitated the development of the reliable, valid, and practical individually administered intelligence test batteries (Neisser et al., 1996) used in the identification of individuals with ID. Space does not allow for a detailed treatment of the lengthy history of research on psychometric-based theories of intelligence in this book. For those interested, more thorough historical accounts can be found in Brody (2000), Carroll (1993), Cudek and MacCallum (2007), Horn and Noll (1997), Schneider and McGrew (2012), and Wasserman (2012).

Psychometric theories of intelligence have their roots in the work of Charles Spearman. Spearman (1904, 1927) initially reported the phenomena of *positive manifold*, or the tendency for all tests of mental ability to be positively correlated. Spearman's *g*-theory posits that all mental tests are positively correlated due to the influence of a common cause, *g* (Jensen, 1998). Although Spearman's *g*-theory is most often described as single factor theory, this characterization is not entirely accurate (Carroll, 1993; Schneider & McGrew, 2012). Scores from individual mental tests were viewed as being due to the influence of *g* and by specific (*s*) abilities unique to each individual test.

From the 1940s to 1960s, psychometric intelligence research focused on the identification of multiple primary abilities. Thurstone's Primary Mental Abilities (PMAs; Thurstone, 1938, 1947; Thurstone & Thurstone, 1941) theory was most prominent and identified between seven and nine PMAs. Thurstone (1947) was willing to accept the possible existence of *g* above the PMAs, but believed it was less important than the PMAs (Carroll, 1993; Schneider & McGrew, 2012). Summaries of this period of intense factor analysis research suggested over 60 possible separate PMAs (Carroll, 1993; Ekstrom, French, & Harman, 1979; French, 1951; French, Ekstrom, & Price, 1963; Guilford, 1967; Hakstian & Cattell, 1974; Horn, 1976).

The next significant phase in the development of psychometric theories of intelligence was driven largely by the research of Raymond Cattell. Based on an extensive program of factor analysis research, Cattell concluded that Spearman's *g* was best explained by splitting it into general fluid (g_f) and general crystallized (g_c) intelligence (Cattell, 1941, 1943). Horn, Cattell, and many others published systematic programs of factor-analytic research (from 1965 to the late 1990s) that confirmed the original Cattell *Gf* – *Gc* model and added new factors. Horn extended the *Gf* – *Gc* theory to eventually include 9–10 broad abilities (Horn, 1989). In 1993, Carroll published his seminal work, *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*, that summarized his re-factor analysis of more than 460 different datasets since the time of Spearman. The result was Carroll's three-stratum hierarchical model of intelligence that included general (*g*) intelligence at the apex, which subsumed eight *broad* intellectual abilities that, in turn, subsumed over 70 *narrow* PMAs. Most other psychometric scholars concur that Carroll's work presented the first validated hierarchical taxonomy of human intelligence (Ackerman & Lohman, 2006; Burns, 1994; Jensen, 2004; Kaufman, 2009; Keith & Reynolds, 2010; McGrew, 2005, 2009; Schneider & McGrew, 2012).

The remarkable similarities between the Carroll and Cattell–Horn models resulted in their integration as the Cattell-Horn-Carroll (CHC) theory of intelligence (Daniel, 1997, 2000; Kaufman, 2009; McGrew, 1997, 2005, 2009; Schneider & McGrew, 2012; Snow, 1998). Eight to nine broad ability domains (fluid intelligence or reasoning, *Gf*; crystallized intelligence or comprehension-knowledge, *Gc*; long-term storage and retrieval, *Glr*; short-term and working memory, *Gsm*; visual-spatial processing, *Gv*;

auditory processing, *Ga;* intellectual processing speed, *Gs;* quantitative knowledge, *Gq*) are generally accepted as the primary foundation of CHC theory (see McGrew, 2005, 2009, and Schneider & McGrew, 2012, for detailed definitions). These five to eight CHC broad ability domains are represented by one or more subtests on some, but not all, intelligence batteries.

Contemporary Neurocognitive, Neuropsychological, and Developmental Research

The content of contemporary intelligence batteries has also been influenced by current theories and research in cognitive neuroscience, neuropsychology, and developmental psychology (Drozdick, Wahlstrom, Zhu & Weiss, 2012; Naglieri, Das & Goldstein, 2012). In particular, recognition of the importance of the constructs of planning, working memory, and intellectual processing speed have resulted in the inclusion of more tests of these abilities in intelligence batteries. The neuropsychological theory of Luria (see Naglieri et al., 2012) has also influenced the revisions of some intelligence batteries. Briefly, Luria's theory defines intelligence as being based on four functional aspects related to brain structures. These four functional components are best articulated in the contemporary Planning, Attention-Arousal, Simultaneous, and Successive (PASS) theory of intelligence, which proposed that cognition consists of the functional brain processes of planning, attention, and simultaneous and successive processing (Naglieri et al., 2012; Singer, Licthenberger, Kaufman, Kaufman, & Kaufman, 2012). Developmental intelligence research, particularly that reporting developmental changes in intellectual processing speed and the dynamic interaction of working memory and processing speed in adulthood, has informed the revision of adult intelligence batteries (Drozdick et al., 2012; Wechsler, 2008).

Available Intelligence Test Batteries

Table 7.1 presents a summary of the comprehensive, nationally normed, individually administered contemporary intelligence batteries that possess satisfactory psychometric characteristics (i.e., national norm samples, adequate reliability and validity for the full-scale IQ score) for use in the diagnosis of ID. Only three of the intelligence batteries (i.e., Stanford-Binet Intelligence Scales, fifth edition [SB5, Roid, 2003], Wechsler Adult Intelligence Scale, fourth edition [WAIS-IV], and Woodcock-Johnson Tests of Cognitive Abilities, third edition [WJ III]) have adult norms suitable for testing adults. However, the files of many adults often contain scores or reports from intelligence testing during the individual's childhood and adolescence. The most commonly administered contemporary childhood and adolescent intelligence batteries are also listed in Table 7.1.

The "full-scale IQ score" column in Table 7.1 lists the full-scale general intelligence score provided by each intelligence battery. This score is the best estimate of a person's general intellectual ability for the purposes of diagnosing ID from each respective battery. All full-scale IQ scores listed in Table 7.1 meet Jensen's (1998) psychometric sampling error criteria for measuring general intelligence (*g*). As recommended by Jensen

TABLE 7.1. Individually Administered Comprehensive Intelligence Batteries

				Component Part		
Intelligence Battery	Publication Date	Age Range (years)	Composite g-score	Name	Scales g loadings[b]	h^2[c]
CAS	1997	5–17	Full Scale (FS)	Simultaneous	0.77	0.59
				Planning	0.75	0.56
				Attention	0.75	0.56
				Successive	0.66	0.44
DAS-II	2007	2–17	General Conceptual Ability (GCA)	Nonverbal Ability	0.81	0.66
				Spatial Ability	0.80	0.64
				Verbal Ability	0.78	0.61
				Working Memory	0.69	0.48
				Processing Speed	0.54	0.29
KABC-II	2004	3–18	Mental Processing Index (MPI) Fluid-Crystallized Index (FCI)	Gf/Planning	0.81	0.66
				Gc/Knowledge	0.81	0.66
				Gv/Simultaneous	0.77	0.59
				Glr/Learning	0.75	0.56
				Gsm/Sequential	0.67	0.45
SB5	2003	2–85+	Full Scale IQ (FS IQ)	Quantitative Reasoning	0.89	0.79
				Knowledge	0.86	0.74
				Visual-Spatial Processing	0.88	0.77
				Fluid Reasoning	0.86	0.74
				Working Memory	0.85	0.72
WAIS-IV	2008	16–90+	Full Scale IQ (FS IQ)	Working Memory Index	0.85	0.72
				Perceptual Reasoning Index	0.85	0.72
				Verbal Comprehension Index	0.83	0.69
				Processing Speed Index	0.74	0.55
WISC-IV	2004	6–16	Full Scale IQ (FS IQ)	Perceptual Reasoning Index	0.84	0.71
				Verbal Comprehension Index	0.83	0.69
				Working Memory Index	0.78	0.61
				Processing Speed Index	0.72	0.52
WJ III /NU	2001, 2007[a]	2–90+	General Intellectual Ability (GIA-Standard; GIA-Extended)	Comprehension-Knowledge	0.74	0.55
				Long-term Storage & Retrieval	0.74	0.55
				Fluid Reasoning	0.70	0.49
				Auditory Processing	0.62	0.38
				Short-term Memory	0.62	0.38
				Visual Processing	0.55	0.30
				Processing Speed	0.52	0.27

Note: CAS = Cognitive Assessment System (Naglieri & Das, 1997); DAS-II = Differential Ability Scales—Second Edition (Elliott, 2007); KABC-II = Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004); SB5 = Stanford Intelligence Scales—Fifth Edition (Roid, 2003); WAIS-IV = Wechsler Adult Intelligence Scale—Fourth Edition (Wechsler, 2008); WISC-V = Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2004); WJ III / NU = Woodcock-Johnson Battery—Third Edition and Normative Udate (Woodcock, McGrew, Schrank, & Mather, 2001, 2007).

a = WJ III was first published in 2001 and then the norms were "freshened" with a normative update in 2007.

b = Within each battery principal component analysis was used to extract a single g-component from correlation matrices reported in the respective technical manuals. Tables A.10 and A.11 were used from the CAS manual, Tables 8.2 and 8.3 from the DAS-II manual; Tables 8.10 to 8.13 from the KABC-II manual, Table A.6 from the SB5 manual, and Tables 5.1 from the WAIS-IV and WISC-IV manuals. For the WJ III the principal axes g-loadings reported by Floyd, McGrew, Barry, Rafael & Rogers (2009) were used. When more than one correlation matrix was analyzed, the median value was calculated and is reported in this column. These values were used to order the respective component part scales from the highest to the lowest within-battery values. It is important to note that these are within-battery estimates and comparisons across the different batteries is not appropriate.

c = communality or percent of variance shared with principal g-factor.

(1998), "the particular collection of tests used to estimate *g* should come as close as possible, with some limited number of tests, to being a representative sample of all types of mental tests, and the various kinds of test should be represented as equally as possible" (p. 85). At a minimum, a measure of general intelligence (i.e., full-scale IQ) should be based on a variety of different tests that vary on information content, skills, and mental operations, and sample from at least three primary intelligence domains (e.g., at least three of the broad CHC intelligence domains) (Jensen, 1998). All IQ test batteries included in Table 7.1 meet these criteria.

Also included in Table 7.1 are the part-scale scores (e.g., WAIS-IV Verbal Comprehension Index, Perceptual Reasoning Index, Working Memory Index, and Processing Speed Index) provided by each battery, followed by their respective within-battery *g*-loadings. This part-score information is included in Table 7.1 as it is relevant to the use of IQ part scores, in place of the full-scale IQ scores, for the diagnosis of ID in certain situations (see the "Use of General Intelligence Full Scale and Composite Part Scores" section for explanation and definition of terms). Space does not permit a detailed discussion and comparison of the strengths and limitations of each of the batteries listed in Table 7.1. Overviews of each of the major intelligence test batteries can be found in Flanagan and Harrison (2012).

Comparability of IQ Scores

"Not all scores obtained on intelligence tests given to the same person will be identical" (Schalock et al., 2010, p. 38). For example, Schalock et al. (2010) reported that, although Wechsler Intelligence Scale for Children, third edition (WISC-III) and Stanford-Binet Intelligence Scales, fourth edition (SB-IV) IQ test scores were significantly correlated in one sample of students with ID at the upper end of the range; on the average, the WISC-III scores were eight IQ points lower. Although the lay public often assumes that IQ scores from different tests should be similar, and for the majority of individuals they are reasonably comparable when using technically sound comprehensive measures of general intelligence, 1-to-1 IQ test score correspondence for all individuals is not supported by the available research. That is, one cannot assume that for all individuals the IQ scores from different IQ tests will be similar—and often they can be markedly different.

Full-scale IQ test scores from different tests are frequently similar or are reasonably close (when measurement error is taken into consideration). In other instances, IQ test scores will be markedly different (Floyd, Clark & Shadish, 2008; Macvaugh & Cunningham, 2009)—a finding that often produces consternation for examiners and recipients of psychological reports. The fundamental issue underlying discussions of IQ-IQ score comparisons in cases is that of IQ battery score exchangeability. "*[E] xchangeability* refers to the assumption that the IQ a person receives will be reasonably constant no matter which intelligence test battery is used" (italics in original; Floyd et al., 2008, p. 415). The obvious differences in the test stimuli, task requirements, and test content among different intelligence test batteries would lead most to the conclusion that not all intelligence scores will be exchangeable:

> Exchangeability is thought to be plausible because of the principle of aggregation . . . [t]hat is, influences associated with individual tests in a battery are averaged out when multiple test scores are aggregated into an IQ. As a result, only a single ability, general intelligence, is thought to remain as the systematic source of variance. (Jensen, 1998; Spearman, 1927; Floyd et al., 2008)

In one of the better investigations of IQ score exchangeability to date, Floyd et al. (2008) evaluated IQ-IQ exchangeability across 10 different IQ battery full-scale IQ scores (comprising 6 to 14 individual tests) across approximately 1,000 subjects from six different IQ test validity study samples. Comparisons included most major individually administered IQ test batteries, such as the Differential Ability Scale (DAS; Elliott, 2007), the Kaufman Assessment Battery For Children—Second Edition (KABC-II, Kaufman & Kaufman, 2004), the Kaufman Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993), the Wechsler Intelligence Scale For Children—Third Edition and Fourth Editon (WISC-III/IV; Wechsler, 1991, 2003) the Wechsler Adult Intelligence Scale—Third Edition (WAIS-III; Wechsler, 1997), and the Woodcock-Johnson Tests of Cognitive Abilities (WJ III; Woodcock, McGrew, & Mather, 2001), in various combinations. Five of the six samples included subjects without disabilities from ages 8 to 16. The sixth sample was a mixed sample of university students with and without learning disabilities. Floyd et al. (2008) reported that different intelligence test batteries produce less similar IQs than expected from the apparent high degree of correlations (r range = .69 to .93; median $r = .76$). These authors concluded that "psychologists can anticipate that 1 in 4 individuals taking an intelligence test battery will receive an IQ more than *10 points higher or lower* [emphasis added] when taking another battery" (Floyd et al., 2008, p. 414).

Similar variability in IQ test scores has been demonstrated in samples of individuals with low general intelligence. Whitaker (2008) conducted a meta-analysis of 18 studies (total $n = 2{,}006$ individuals) that investigated the stability of IQ scores for individuals with IQ test scores less than 80. Across the various editions of the Wechsler scales (i.e., Wechsler Intelligence Scale for Children—WISC; Wechsler Intelligence Scale for Children—Revised, WISC-R; WISC-III; Wechsler Adult Intelligence Scale, WAIS; Wechsler Adult Intelligence Scale—Revised,WAIS-R) and the 1960 and 1972 Stanford-Binet, the average stability coefficient (mean assessment interval of 33.5 months) for the total IQ test score was .82. In the eleven studies with adequate information, 57 % of the IQ test scores changed by less than six IQ test score points. However, 14 % of the individuals had IQ test scores that changed by 10 points or more. Given the bright line emphasis in many *Atkins* cases, a shift of one IQ test score point (e.g., from 70 to 71) can often result in a different decision by the courts. A change of 10 IQ test score points could move an IQ test score that is clearly within the ID score range (e.g., 68) to a score (e.g., 78) that is beyond the bright line cutoff score of 70 and even beyond the upper bound score (e.g., 75) that accounts for the standard error of measurement (± 5 IQ test score points).

IQ-IQ Score Differences: Basic Measurement Concepts

Understanding why different full-scale IQ scores may be reported for an individual on different occasions or from different intelligence batteries requires an understanding of a number of basic measurement concepts. This discussion builds on the topics addressed in Chapter 5, which provides a detailed discussion of basic measurement concepts.

Intelligence score correlations. Comparing different IQ scores requires an understanding of the statistical concept of correlation. The *APA Dictionary of Psychology* (VandenBos, 2007) defines the related statistical concepts of *correlation, correlation coefficient*, and the *coefficient of determination* as:

> *Correlation:* "The degree of relationship (usually linear) between two attributes." (p. 234)
>
> *Correlation coefficient (r):* "A numerical index reflecting the degree of relationship (usually linear) between two attributes scaled so that the value of +1 indicates a perfect positive relationship, −1 a perfect negative relationship, and 0 no relationship." (p. 234)
>
> *Coefficient of determination (r^2).* "A numerical index that reflects the degree to which variation in the dependent variable is accounted for by one independent variable. Also called **determination coefficient.**" (bold emphasis in original; p. 186)

Correlations reported between full-scale IQ scores from the major individually administered intelligence batteries usually range from the .60s to .80s, with the highest correlations reported in the .70 to .80 range. Although these are statistically significant high correlations, it is important to recognize that correlations estimate the relations of two IQs across individuals (i.e., in the research sample group) and can lead to a false sense of expected IQ-IQ correspondence for a specific individual.

The *coefficient of determination* (r^2) is most informative in understanding expected score similarities or differences as it quantifies the amount of shared or common test score variance between the two tests (Neisser et al., 1996; Sattler, 2001). This index is obtained by squaring a reported correlation (e.g., .70 × .70 = .49) and multiplying the result by 100. For example, if $r = .70$, the result is 49.0%. What does this statistic mean? In this example, the .70 correlation indicates that the global IQ scores from two different test batteries have approximately 50% common or shared test score variance. The remaining 50% of unshared variance is due to (a) different abilities being measured by the two different intelligence test batteries; and (b) to a lesser extent, measurement error due to less than perfect reliability for each test score. Knowing that two intelligence test batteries may have approximately 50% shared (common) and unshared (uncommon) IQ test score variance should lead the reader to the conclusion that not all individuals will receive the same IQ test score (or nearly similar scores) on two different intelligence tests that correlate at .70.

An example is provided to illustrate this important point. In the third edition of WAIS's (WAIS-III) technical manual (Wechsler, 1997) a correlation of .88 (statistically significant and high) is reported between the WAIS-III Full Scale IQ and the Stanford-Binet Intelligence Scale–Fourth Edition (SB4) global score ($n = 26$ adult subjects). A correlation of .65 is also reported between the WAIS-III IQ and the special purpose Raven's Standard Progressive Matrices (SPM; Raven, 1976) in the same sample. Correlations of this magnitude, when converted to coefficients of determination, indicate that the WAIS-III has approximately 77% and 42% common or shared variance with the SB4 and SPM, respectively. The WAIS-III/SB4 77% value is high and impressive. Yet, again, it is important to recognize that this group study suggests that the WAIS-III and SB4 still have 23% (approximately 1/4) of their respective test score variance that they do *not* share in common. The WAIS-III and SPM have more they do not share (58%) than they do measure in common (42%).

The only time one can expect two different intelligence tests to provide approximately the same IQ test scores for all individuals is if the tests are nearly perfectly correlated (correlation approaches +1.0). This is not the reality reflected by decades of IQ test comparison research. Although the typical correlations reported between major intelligence tests (.60s to .80s) may sound impressive to nonpsychometricians, correlations of this magnitude suggest that different intelligence tests measure approximately 40% to 60% common abilities and, thus, different IQs are to be expected with regularity (Floyd et al., 2008).

Before interpreting differences between two IQ test scores, one must first determine if the IQ-IQ difference score is a statistically significant and reliable difference. That is, one must determine if the IQ-IQ score difference is not simply due to chance. Understanding a number of statistical concepts is necessary to make this evaluation—*reliability* and the *standard error of the difference score* are briefly defined below (see Chapter 5 for a more thorough discussion).

> *Reliability:* "The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group." (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999, p. 180) Reliability "refers to the consistency of measurements." (Sattler, 2001, p. 102)
>
> *Reliability coefficient* (r_{11})*:* "Expresses the degree of consistency in the measurement of test scores. The symbol is the letter *r* with two identical subscripts (r_{xx} or r_{tt}). Reliability coefficients range from 1.00 (indicating perfect reliability) to .00 (indicating the absence of reliability." (Sattler, 2001, p. 102)
>
> *Standard error of difference score* (SE_{diff})*:* "This statistic provides an estimate for the standard deviation of the sampling distribution of the difference between the

> two obtained index scores. Multiplying the SE_{diff} by an appropriate z value yields the amount of difference required for statistical significance at any given level of confidence." (italics in original; Wechsler, 2008, p. 53)

It is important to note that the reliability of the difference between two IQ test scores (IQ-IQ = IQ difference score) will be smaller than the reliability of the two individual scores. The *standard error of difference score* (SE_{diff}; Anastasi & Urbina, 1997) reflects this statistical fact by incorporating the reliability of the two compared scores in the calculation of the SE_{diff}, which is then used to evaluate the statistical significance of an IQ-IQ difference score. It is important to note that the SE_{diff} *for* a difference between two scores will be larger than the standard error of measurement (*SEM*) of the individual scores. Reliabilities for composite IQ scores should be available in each test's technical manual. When the tests are on the same standard scale ($M = 100$; $SD = 15$) the formula for calculating the SE_{diff} using the respective reliability coefficients for each IQ score (r_{11} and r_{22}), is:

$$SE_{diff} = 15 \times \sqrt{[2 - r_{11} - r_{22}]}$$

Given two IQ tests with full-scale IQ score reliabilities of .95 and .93, $SE_{diff} = 15 \times \sqrt{[2 - .95 - .93]}$, or $SE_{diff} = 5.2$. To determine how large a score difference could be obtained by chance (.05 level of probability), SE_{diff} (5.2 in example) is multiplied by 1.96 with a result of approximately 10.2 (rounded to 10 for discussion purposes). In this example, before interpreting the differences between two IQ test scores as significant ($p < .05$), an IQ-IQ difference would need to be at least ±10 points.

It is important to note that the above example is based on internal consistency derived reliability for each of the two test batteries and does not incorporate the lower test-test reliability (stability) present when comparing IQ test scores across time. For example, in the previously mentioned Whitaker (2008) meta-analysis of the stability of IQ test scores for individuals with low general intelligence, the 95% IQ-IQ test score difference confidence interval may be as large as approximately ±12.5 IQ points for two IQ test scores separated, on average, by approximately three years. (See Chapter 5 for further discussion of stability of IQ scores across time.)

Possible Explanations for IQ-IQ Score Differences

Factors contributing to significant IQ score differences are many and may include: (a) procedural or test administration errors (e.g., scoring errors, improper nonstandardized test administration, malingering, age vs. grade norms, practice effects); (b) test norm or standardization differences (e.g., norm obsolescence or the Flynn effect); (c) content differences across different test batteries or between different editions of the same battery; or (d) variations in a person's performance on different occasions (Floyd et al., 2008; McGrew, 1994; Schalock et al., 2010). Due to space limitations, only parts of (a) and (c) will be addressed here, as Chapters 5 and 10 address issues involved in topics (b) and (d).

Test procedural and administration errors. Ramos, Alfonso, and Schermerhorn (2009) summarized the extant research on examiner errors and reported sufficient average examiner error to produce significant changes in IQ test scores for many individuals. The most frequent types of errors reported included a failure to record responses, use of incorrect basal and ceiling rules, reporting an incorrect global IQ test score, incorrect adding of subtest scores, incorrect assignment of points for specific items, and incorrect calculation of the individual's age. On Wechsler-related studies, Ramos et al. (2009) found that studies have reported average error rates from 7.8 to 25.8 errors per test record, almost 90% of examiners making one error, and, in one study, two thirds of the reviewed test records resulted in a change in the full-scale IQ. Examiner errors do not appear to be instrument-specific, as Ramos et al.'s study reported an average error rate of 4.6 errors per test record on the WJ III.

The importance of verifying accurate administration and scoring is evident in the finding that, across both experienced psychologists and students in graduate training, differences between original obtained IQ scores and correctly scored IQ scores were as high as 25, 22, and 22 points for the WAIS-III Verbal, Performance, and Full Scale IQ test scores, respectively (Ryan & Schnakenberg-Ott, 2003). Despite examiners reporting confidence in their scoring accuracy, Ryan and Schnakenberg-Ott reported average levels of agreement with the standard (accurate) test record of only 26.3% (Verbal IQ), 36.8% (Performance IQ), and 42.1% (Full Scale IQ). This level of examiner error is alarming, particularly in the context of IQ test score-based life-and-death decisions such as in *Atkins* cases.

Content differences between IQ test batteries and within different editions of the same IQ test battery. As is often the case in *Atkins* cases, individuals have frequently been tested multiple times and often with different editions of a battery (e.g., WAIS-R, WAIS-III, and WAIS-IV). Psychologists who compare and interpret the consistency or variability of these scores must be aware of significant content changes across editions that may explain differences in the full-scale IQ scores. This point is illustrated with the adult Wechsler battery in Table 7.2. The points made also pertain to changes in different versions of other intelligence batteries and are not specific to the adult Wechsler series.

First, it is important to know that the adult Wechsler scales are based on 10 or more subtests that are added together to provide part scores (e.g., Verbal IQ, Performance IQ), as well as the full-scale IQ test score. The original Wechsler Adult Intelligence Scale (WAIS) and its second, revised, version (WAIS-R) were both comprised of six verbal tests that produced a Verbal IQ and five nonverbal tests that produced the Performance IQ score. The eleven tests together comprised the WAIS and WAIS-R Full Scale IQ. This is illustrated by the first two columns in Table 7.2. The gray shading indicates that the same eleven subtests were the basis of the WAIS and WAIS-R Full Scale IQ scores.

TABLE 7.2. Changes in Subtests Contributing to Adult Wechsler Full Scale IQ Scores Across Four Editions

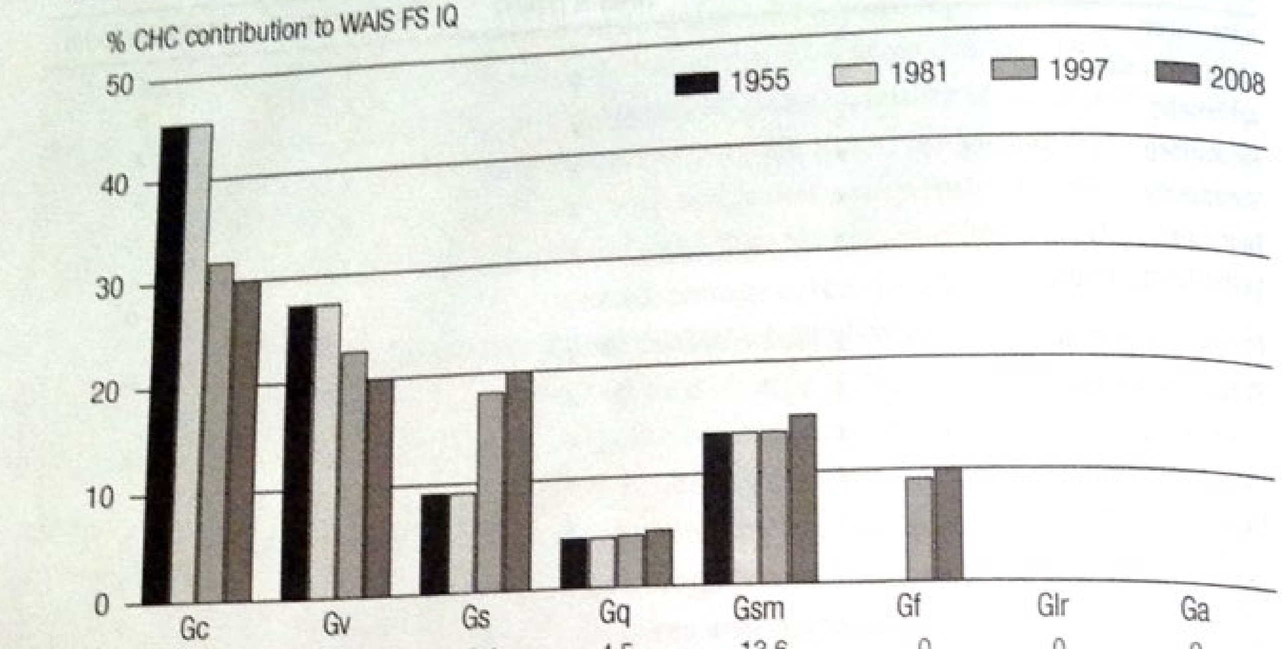
Subtests	*WAIS (1955)*	*WAIS-R (1981)*	*WAIS-III (1997)*	*WAIS-IV (2008)*
Information	X	X	X	X
Comprehension	X	X	X	0
Arithmetic	X	X	X	X
Similarities	X	X	X	X
Vocabulary	X	X	X	X
Digit Span	X	X	X	X
Letter-Number Seq.			0	0
Picture Completion	X	X	X	0
Picture Arrangment	X	X	X	
Block Design	X	X	X	X
Object Assembly	X	X	0	
Digit Sym/Coding	X	X	X	X
Symbol Search			0	X
Matrix Reasoning			X	X
Visual Puzzles				X
Figure Weights				0
Cancellation				0

Note. X = subtests included in the calculation of the FS IQ for each edition of the WAIS. 0 = Supplemental tests. Shading demonstrates continuity of comosition of FS IQ scores.

As can be seen in Table 7.2, the WAIS-III started a process of revision to the adult Wechsler's wherein which all original 11 subtests were retained, but new subtests were added. More important is the fact that those WAIS-III subtests, designated by gray shading, no longer match the exact set of 11 subtests as in the WAIS and WAIS-R. Thus, the WAIS-III Full Scale IQ test score is based on a different mixture of subtests and abilities than the earlier WAIS and WAIS-R. When the WAIS-IV was published, it contained 15 subtests. More importantly, those subtests that contributed to the Full Scale IQ are not 100% comparable to the same set as in the WAIS-III or the WAIS and WAIS-R. The important conclusion from Table 7.2 is that, as the adult Wechsler battery evolved, the specific combination of tests that comprise the Full Scale IQ (i.e., the score used to aid in the diagnosis of ID) changed in composition. The result is that the full-scale IQ scores from the later WAIS-III and WAIS-IV are not 100% comparable in subtest (and abilities measured) to the earlier WAIS and WAIS-R, nor are the latest editions 100% comparable.

The differences in the ability domains measured by different intelligence batteries can produce significant and valid IQ score differences. Furthermore, when a specific line of intelligence test batteries is revised (e.g., Wechsler-Bellevue Intelligence Scale, WAIS, WAIS-R, WAIS-III, WAIS-IV; see Table 7.2), changes are often made to item content and old subtests are eliminated, demoted to supplemental status, or are

FIGURE 7.1. Changes in Proportional CHC Abilities Represented in Adult Wechsler Full Scale IQ Score Across Editions



	Gc	Gv	Gs	Gq	Gsm	Gf	Glr	Ga
1995	45.5	27.3	9.1	4.5	13.6	0	0	0
1981	45.5	27.3	9.1	4.5	13.6	0	0	0
1997	31.8	22.7	18.2	4.6	13.6	9.1	0	0
2008	30	20	20	5	15	10	0	0

Source: McGrew, K. (2010). *Applied psychometrics 101 brief #6: Understanding Wechsler IQ score differences—the CHC evolution of the Wechsler FS IQ score.* Retrieved from http://www.iqscorner.com/2010/02/ap101-brief-6-understanding-wechsler-iq.html

Gc = Comprehension-knowledge; Gv = Visual-spatial processing; Gs = intellectual processing speed; Gq = quantitative knowledge; Gsm = short-term and working memory; Gf = Fluid intelligence or reasoning; Glr = long-term storage and retrieval; Ga = auditory processing.

replaced with completely new subtests. These changes can result in different editions of the similarly named intelligence battery (e.g., WAIS-R, WAIS-III, WAIS-IV) providing full-scale IQ test scores with enough substantive content differences to change the composition or flavor of the total IQ test scores that are compared.

Figure 7.1 demonstrates the changes in the abilities represented by the full-scale IQ test scores from the different editions of the adult Wechsler battery when results from CHC-based within- and cross-battery factor analysis studies have been completed (see Flanagan, Ortiz, & Alfonso, 2007; Keith & Reynolds, 2010; McGrew, 1997; McGrew & Flanagan, 1998, for summaries of this research). It is clear that the full-scale IQ scores from the 1955 WAIS and 1981 WAIS-R were measuring similar abilities. However, the advent of the 1997 WAIS-III resulted in a shift in abilities measured—less *Gc* (verbal) and *Gv* (visual-spatial), more *Gs* (processing speed), but similar proportions of *Gq* (quantitative knowledge) and *Gsm* (short-term and working memory). The decreasing importance of *Gc* and *Gv* continued in the 2008 WAIS-IV concurrently with a slight increase in *Gs* and *Gsm*. *Gq* and *Gf* were approximately the same in proportional contribution to the WAIS-IV Full Scale IQ as was in the WAIS-III. It should be obvious from this review that different IQ scores may result from individuals who have taken different versions of the adult Wechsler batteries. An understanding of the abilities comprising the composite IQ test score in intelligence batteries is required when attempting to understand and interpret possible IQ-IQ score differences within and across different series of intelligence tests.

It is clear from the above discussion that understanding a range of technical issues may be required when dealing with *Atkins* cases where significant IQ-IQ test score variability is present. If the professionals who administer or interpret the scores from IQ tests do not possess the necessary expertise regarding these technical issues and literature, then consultation with specialists who possess such expertise should be considered.

Use of General Intelligence Full Scale and Composite Part Scores

Examiners are typically faced with IQ battery subtest or part-score profiles that display some degree of variability between the part scores (e.g., the four WAIS-IV index scores) or between the individual subtest scores. The extant research suggests that a certain degree of within-profile variability is normal and not diagnostically significant. However, there are situations when the observed score variability is so large that the validity of the total full-scale IQ should be questioned. More importantly, there are situations where select IQ battery component part scores may be better estimates of an individual's general intelligence than the full-scale IQ. As summarized by Reschly, Meyers, & Hartel,

> whenever the validity of one or more part scores (subtests, scales) is questioned, examiners must also question whether the test's total score is appropriate for guiding diagnostic decision making. The total test score is usually considered the best estimate of a client's overall intellectual functioning. However, there are instances in which, and individuals for whom, the total test score may not be the best representation of overall cognitive functioning. (2002, p. 106–107).

In the Reschly et al. (2002) National Academy of Sciences report, "Committee member Keith Widaman dissents from this part of the recommendation. Dr. Widaman believes that IQ part scores representing crystallized intelligence (*Gc*, similar to Verbal IQ) and fluid intelligence (*Gf*, related to performance IQ) have clear discriminant validity and represent broad, general domains of intellectual functioning" (Reschly et al., 2002, p. 3, footnote 1).

A number of issues must first be considered before using component part scores to estimate an individual's level of general intelligence—statistical significance, meaningful differences, and appropriate intellectual abilities (Reschly et al., 2002). First, preference for part scores (instead of the full-scale IQ) should occur only if there are statistically significant differences between the intelligence battery part scores that contribute to the full-scale IQ. The technical manual and/or computer-generated score interpretation report for each intelligence battery typically includes the necessary information to determine if the part scores are statistically different. If not, "the total score is unequivocally the best indicator of overall cognitive functioning and should be used for decision making" (Reschly et al., 2002). Second, the presence of statistically significant part score differences is a necessary, but not sufficient, condition for not interpreting the full-scale IQ score. The differences must also be meaningful differences—relatively rare or unusual in the general population. This is often referred to as base rate, which is defined as "the unconditional, naturally occurring rate of a phenomenon in a population" (VandenBos, 2007, p. 103). Statistically significant part-score differences that occur less frequently than approximately 25% of the general population have been recommended as an appropriate base rate for considering IQ part scores in the determination of general intelligence (Reschly et al., 2002). Collectively, these first two points indicate that part-score differences should be *both* statistically significant and relatively unusual in the population before an examiner determines that the total full-scale score is not the best indicator of an individual's general level of intellectual functioning. If these two conditions are met, then one or more of the part scores might be used to estimate the individual's general intelligence. However, the examiner cannot just use any part score(s) for ID determination. Only the most appropriate part-score measures should be used to diagnose ID. The use of part scores is not the generally accepted method for diagnosing ID and should only be used in unusual cases where the validity of the full-scale IQ score is clearly in doubt. The use of part scores in idiosyncratic "junk science" interpretations is not appropriate.. The joint test standards established by AERA, APA, & NCME (1999), in particular, should be adhered to in these unusual cases and would include, at a minimum, (a) providing evidence to support the use of particular part scores as the best proxies for estimating general intelligence for the specific case (Standard 1.4); (b) minimizing potential misinterpretations and unintended consequences in the use of part scores (Standard 11.5); (c) the articulation of a scientific-based logical analysis of relevant reliability and validity evidence to support inferences and interpretations (Standard 12.13); and (d) use of multiple sources of convergent and collateral data to support the unique case-specific interpretation (Standards 12.18 and 12.19). The use of part scores should not be used as a justification for abbreviated evaluations, a means to not sufficiently explore diagnostic questions, or to solve sociopolitical problems. Formalized clinical judgment, as articulated by Schalock and Luckasson (2005), must be followed. Such clinical judgment is characterized as "being systematic (i.e., organized, sequential, and logical), formal (i.e., explicit and reasoned), and transparent (i.e., apparent and communicated clearly)" (p. 1).

The issue of appropriate intellectual abilities deals with which part scores within an intelligence battery are most associated with general intelligence (*g*). The part scores that are more associated with general intelligence are often referred to as the high *g*-loading scores. The research- and theory-based consensus is that measures of *Gc* and *Gf* are the highest *g*-loading measures and constructs and are the most likely candidates for elevated status in diagnosing ID (Reschly et al., 2002). However, examination of the *g*-ness of composite scores from existing batteries (last three columns in

Table 7.1) suggests this traditional assumption may *not* hold across all intelligence batteries. (The h^2 values are the values that should be used to compare the relative amount of *g*-variance present in the component part scores within each intelligence battery.)

In the case of the SB5, all five composite part scores are very similar in *g*-loadings ($h^2 = .72$ to .79). No single SB5 composite part score appears more superior to the other scores when attempting to diagnose ID on the basis of these scores (and not the full-scale IQ score). At the other extreme is the WJ III, where the Fluid Reasoning, Comprehension-Knowledge, and Long-Term Storage and Retrieval clusters scores are the best *g*-proxies for part-score based interpretation. The WJ III Visual Processing and Processing Speed clusters are not composite part scores that should form the primary basis for an argument of ID due to their relatively low *g*-loadings. Across all batteries that include a processing speed component part score (i.e., Differential Ability Scales–Second edition [DAS-II], WAIS-IV, WISC-IV, WJ III), the processing speed scale is always the weakest proxy for general intelligence and, thus, would not be viewed as a good standalone estimate of general intelligence.

It is also clear that one cannot assume that composites with similar sounding names of measured abilities will have similar relative *g*-ness status within different batteries. For example, the *Gv* (visual-spatial or visual processing) clusters in the DAS-II (Spatial Ability) and SB5 (Visual-Spatial Processing) are relatively strong *g*-measures within their respective battery, but the same cannot be said for the WJ III Visual Processing cluster. Even more interesting are the differences in the WAIS-IV and WISC-IV relative *g*-loadings for similarly sounding index scores.

For example, the Working Memory Index is the highest *g*-loading component part score (tied with Perceptual Reasoning Index) in the WAIS-IV, but is only third (out of four) in the WISC-IV. The Working Memory Index comprises the Digit Span and Arithmetic subtests in the WAIS-IV and the Digit Span and the Letter-Number Sequencing subtests in the WISC-IV. The Arithmetic subtest has been reported to be a factorially complex test that may tap fluid intelligence (quantitative reasoning), quantitative knowledge, working memory, and possibly processing speed (Keith & Reynolds, 2010; Phelps, McGrew, Knopik & Ford, 2005). The factorially complex characteristics of the Arithmetic subtest (which, in essence, makes it function like a mini-*g* proxy) would explain why the WAIS-IV Working Memory Index is a good proxy for *g* in the WAIS-IV, but not the WISC-IV.

The above within and across intelligence battery examples of relative part score *g*-ness illustrate that those who pursue a diagnosis of ID based on such scores must be aware of the composition and psychometric *g*-ness of the component scores of the intelligence battery scores interpreted. This is not a new problem in the context of naming factors in factor analysis and, by extension, factor-based intelligence test composite scores. Cliff (1983) described this nominalistic fallacy in simple language—"if we name something, this does not mean we understand it" (p. 120). Not all component part scores in different intelligence batteries are created equal (with regard to *g*-ness).

Finally, before one or more part scores are used to estimate a person's general level of intellectual functioning (in place of the full-scale IQ score), the part score(s) should be evaluated to determine that it is representing a unitary ability. In the context of the WAIS-IV, but relevant to all part scores within all intelligence batteries, Lichtenberger and Kaufman (2009) explained that

> a unitary ability is an ability (such as Crystallized Intelligence or Processing Speed) that is represented by a cohesive set of scale scores, each reflecting slightly different or unique aspects of the ability. Thus, when the variability among the subtest scale scores that compose a WAIS-IV Index is not unusually large, the ability presumed to underlie the index is considered unitary and may be interpreted. (p. 167)

The technical manuals and/or scoring interpretative software for intelligence batteries typically provide the necessary information that allows examiners to ascertain if the variability between the subtests that comprise a part score is relatively consistent and, thus, indicating that a part score can be interpreted as a measure of a valid intellectual ability. If significant and meaningful differences are present among the subtest scores for a part score, then the part score may not be interpretable (Lichtenberger & Kaufman, 2009).

Relation Between Intelligence and Adaptive Behavior Scores

Given the pivotal role intelligence tests and scales of adaptive behavior (AB) play in the diagnosis of ID, it is important to know the typical relation (correlation) between their respective scores. Numerous AB/IQ correlations studies were published in the late 1970s and 1980s between a wide variety of adaptive behavior scales and intelligence tests. Probably the best synthesis of this research was provided by Harrison (1987), which included a table of over 40+ AB/IQ correlations. Harrison (1987) concluded that "the majority of correlations fall in the moderate range" (p. 39). When the correlations with maladaptive measures are excluded from Harrison's table, the correlations range from .03 to .91. Harrison could not identify a specific explanation for the variability or range of the correlations.

The Committee on Disability Determination for Mental Retardation published a National Research Council report (*Mental Retardation: Determining Eligibility for Social Security Benefits*; Reschly et al., 2002) that also addressed the AB/IQ relationship. The report concluded that AB/IQ studies report correlations

> ranging from 0 (indicating no relationship) to almost +1 (indicating a perfect relationship). Data also suggest that the relationship between IQ and adaptive behavior varies significantly by age and levels of retardation, being strongest in the severe and moderate ranges and weakest in the mild range. There is a dearth

> of data on the relationship of IQ and adaptive behavior functioning at the mild level of retardation. (p. 8)

Factors identified as moderating the AB/IQ correlation were scale content, measurement of competences versus perceptions, sample variability, ceiling and floor problems of the scales, and level ofintellectual disability .

Recently, McGrew (2012) combined the 40+ AB/IQ correlations from Harrison (1987) with those reported in the technical manuals of the three most frequently used contemporary adaptive behavior scales (i.e., Vineland Adaptive Behavior Scale, Sparrow, Cicchetti & Balla, 2005; Adaptive Behavior Scales—II, Harrison & Oakland, 2008; Scales of Independent Behavior-Revised, Bruininks, Woodcock, Weatherman & Hill, 1996). Also, the latent AB/IQ correlations (as estimated from confirmatory factor analysis models) reported by Ittenbach, Spiegel, McGrew, and Bruininks (1992); Keith, Fehrmann, Harrison, and Pottebaum (1987); and McGrew and Bruininks (1990) were included. This resulted in the addition of 17 AB/IQ correlations to the 43 from Harrison, for a total of 60 correlations. Focus was only on the composite IQ and AB correlations and not the part scores from the respective measurement instruments.

The 60 AB/IQ correlations ranged from .12 to .90 with a mean of .51, a median of .48, and a standard deviation of .20. McGrew (2012) concluded that an estimate of the typical AB/IQ correlation is approximately .50, with most correlations ranging from approximately .40 to .65. This finding is consistent with Harrison's (1987) conclusion of a moderate correlation. In practical terms, this means that, for any individual, standard scores from AB and IQ scales will frequently diverge and will not always be consistent.

Harrison (1987) provides a succinct explanation for the primary reasons for the moderate correlation between AB and IQ:

> Although intelligence and adaptive behavior scales have many similarities in purposes and uses, several basic differences in the two types of scales warrant this type investigation. According to Meyers et al. (1979), the measurement of intelligence and adaptive behavior differs in several respects, including the following: (1) intelligence scales emphasize thought processes while adaptive behavior scales emphasize everyday behavior, (2) intelligence scales measure maximum performance or potential while adaptive behavior scales measure typical performance, and (3) intelligence scales presume a stability in scores while adaptive behavior scales presume modifiability of performance. (p. 39)

Implications for Practice

The following implications for practice are based on the integration of the content of the current chapter and recommendations from Schalock et al. (2010), the Committee on Disability Determination for Mental Retardation (Reschly et al., 2002), and the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). In addition, selected ethical and professional practice guideline publications were also considered in identifying implications (Gold et al., 2008; Macvaugh & Cunningham, 2009).

Intelligence test battery selection. The diagnosis of ID should be based on an individually administered, comprehensive, nationally, and recently normed intelligence battery that yields a full-scale score of general intellectual functioning (*g*-composite score). Intelligence batteries used in the diagnosis of ID must meet the appropriate professionally accepted psychometric standards of reliability, validity, representative norm sample, and test fairness for diagnostic purposes as per the joint test standards (AERA, APA, & NCME, 1999; see Reschly et al., 2002). The intelligence batteries listed in Table 7.1 meet this definition and standards and represent the most likely comprehensive, individually administered intelligence test batteries found in the records of *Atkins* cases. (The listing of the intelligence batteries in Table 7.1 indicates they meet the primary psychometric criteria for providing a valid measure of general intelligence in a representative sample. However, the publication dates of each battery should be inspected, as some of the batteries may have significant norm obsolescence (i.e., Flynn effect) issues.) Furthermore, often in *Atkins* cases, the individual has a file that may contain historical reports of prior intelligence test scores. In general, a majority of the earlier predecessors of the batteries listed in Table 7.1 meet most psychometric criteria for aiding in ID determination. There are too many exceptions to this statement to be covered in this chapter. For example, as described earlier with regard to the evolution of the WAIS to the WAIS-IV, the newer versions (WAIS-III and WAIS-IV) have benefited from decades of research on intelligence and are better measures of general intelligence (more aligned with contemporary research and theory) than the earlier WAIS and WAIS-R. The same applies to the other series of intelligence batteries listed in Table 7.1. Short-form, group, or computer-administered tests are inappropriate for determining an individual's level of general intellectual functioning for diagnostic purposes. Less comprehensive special purpose intelligence measures (e.g., measures that use nonverbal test administration procedures) should only play a role in estimating general intelligence when individual-specific characteristics (e.g., test fairness issues due to cultural, social, ethnic, or language factors) clearly suggest that the comprehensive individual intelligence batteries are inappropriate for this purpose. Several qualifying considerations are also discussed as follows:

Comprehensive intelligence battery defined. A measure of general intelligence (i.e., full-scale IQ) should be based on a variety of different tests that vary on information content, skills, and mental operations, and sample from at least three primary intelligence domains (e.g., at least three of the broad CHC intelligence domains; Jensen, 1998). These criteria ensure that the general intelligence estimate for an individual is based on a multidimensional (versus a unidimensional) estimate of intelligence. All intelligence batteries listed in Table 7.1 meet these criteria. Given the assumptions of exchangeability, the full-scale IQ scores from these batteries can be thought of as largely interchangeable, although there will be specific situations and unique characteristics of examinees for which the assumption will be untenable (see "Comparability/Exchangeability of IQ Recommendation" section).

Credentials of the individual who is undertaking the assessment. The diagnosis of ID is a serious task, particularly in the context of *Atkins* cases. *Atkins*-related intellectual assessments should be completed and interpreted only by examiners who: (a) have experience with individuals who have ID, (b) are qualified in terms of professional and state regulations and licensing laws, (c) have met the test publisher's guidelines for using a specific battery, (c) are familiar with the strengths and limitations of the intelligence batteries from which they interpret full-scale IQ test scores, (d) are familiar with the assumptions and recommendations of the current AAIDD manual and relevant professional guidelines and principles of practice , and (e) are familiar with all the joint test standards relevant to the use of psychological tests for diagnostic and classification purposes (see also Chapter 22).

Measurement error and cut-scores. Intelligence tests are fallible instruments that result in the obtained full-scale IQ test score potentially being influenced by a number of sources of unreliability or measurement error (e.g., examinee characteristics, examiner influences, environmental conditions, psychometric issues; see Reschly et al., 2002 for discussion). The full-scale IQ test score must be reported with a 95% confidence interval based on the test's standard error of measurement (*SEM*), which for almost all intelligence batteries is approximately ±5 IQ test score points (see Chapter 5). The use of the *SEM* ensures that a specific IQ test score is interpreted as existing within a range of scores reflecting the known measurement error of the intelligence battery. The obtained score is only an estimate of an individual's "true" score. Examiners who provide reports, depositions, or expert testimony have an ethical and professional responsibility to educate the recipient(s) of their interpretations that the diagnosis of ID should not hinge on a single "bright line" specific cut-score. Rigid point-specific IQ cut-scores are arbitrary and fail to reflect the complexities of proper diagnosis of ID, especially a mild level of ID. Specific obtained scores must be interpreted within the context of the 95% confidence interval based on the *SEM*.

Test administration and scoring errors. The frequency and magnitude of test administration and scoring errors reported in intelligence testing research warrants special comment and recommendation. Before attempting to interpret any IQ test scores in *Atkins* cases, or trying to reconcile IQ-IQ test score differences between tests, the first step should be to seek an independent review of the examinee's test records. Any obvious errors should be corrected and new scores recalculated if necessary. Only then should professionals proceed to draw conclusions about scores. If possible, any intelligence test results used in an *Atkins* hearing should be subject to independent review of the original test protocol to ensure against administration or scoring errors that might result in significant differences in the reported IQ test score. Several professional training and monitoring recommendations have empirically demonstrated improvement in reducing such errors (see Erodi, Richard, & Hopwood, 2009; Hopwood

& Richard, 2005; Kuentzel, Hetterscheidt, & Barnett, 2011; Ramos et al., 2009; Ryan & Schnakenberg-Ott, 2003).

The Flynn effect. The Flynn effect (FE) relates to norm obsolescence, which must be recognized and incorporated in the interpretation of IQ test scores when scores from older editions of intelligence batteries are reported. In cases where a test with out-of-date norms was used, a correction for the age of the norms is warranted (Schalock et al., 2010; Schalock et al., 2012). See Chapter 10 for a thorough discussion of norm obsolescence.

Comparability and exchangeability of IQ scores. Succinctly stated, "not all scores obtained on intelligence tests given to the same person will be identical" (Schalock et al., 2010, p. 38). Professionals reporting multiple current or historical IQ scores must be aware of and make appropriate interpretations that recognize the various factors that may contribute to significant IQ test score differences. Assessment professionals should:

a. *Recognize the changing content of contemporary IQ test batteries.* The fact that contemporary intelligence test batteries have become more multidimensional as per contemporary intelligence research and theory warrants special comment and recommendations. Intelligence test battery content differences may be one of the more salient explanations of significant IQ-IQ test score differences. Assessment professionals interpreting IQ-IQ test score differences must integrate knowledge of the changing nature of the full-scale IQ test scores between different contemporary intelligence test batteries (e.g., WAIS-IV vs. SB5; see Flanagan & Harrison, 2012; Flanagan et al., 2007; and Keith & Reynolds, 2010), as well as content differences between older and newer versions of the same battery (e.g., WAIS-R vs. WAIS-III; WAIS-R vs. WAIS-IV), to appropriately interpret the score differences and to render an appropriate professional opinion regarding an individual's general level of intellectual functioning. If an assessment professional integrates the results of historical intelligence testing, it is the professional's responsibility to be familiar with the psychometric characteristics of these tests and possible content changes between instruments so that current and historical testing can be integrated in a scientifically based, professional, and ethical manner in the context of accepted clinical judgment procedures (Schalock & Luckasson, 2005). The *Mental Measurement Yearbook* online service (http://buros.unl.edu/buros/jsp/search.jsp) is accessible to assessment professionals when historical test information is needed.

b. *Prior to interpreting IQ-IQ test score differences, the SE_{diff} statistic should be used to determine if IQ-IQ test score differences are significant.* It is recommended that the discussion and interpretation of discrepancies between IQ test scores incorporate the SE_{diff} statistic to evaluate whether a difference is statistically significant.

c. *Interpretation of multiple and discrepant IQ scores.* When reports contain multiple IQ test scores, either given concurrently or over the lifespan of an individual, examiners should ascertain (to the best of their abilities given the psychological reports available) whether the respective full-scale scores that are compared were judged to be reliable and valid estimates at the original time of testing. When multiple reliable and valid IQ test scores are available, the goal is not to identify which single score is the "best" estimate of an individual's general intelligence. Assessment professionals should integrate the multiple scores and provide a scientific and professionally accepted estimate, using reliable and valid principles and methods, of the person's general level of intellectual functioning. When the multiple scores are reasonably consistent (i.e., a convergence of indicators) and any significant differences are explainable, assessment professionals can have greater confidence in their diagnostic conclusion. Conversely, when major score differences are present in a collection of IQ test scores, interpretations require assessment professionals to educate the recipients of their findings regarding the potential reasons for the IQ test score variability. For example, assessment professionals should address issues such as practice effects, stability of intelligence over time, content differences between different batteries, the Flynn effect, and other issues that affect the comparability of IQ scores included in their written reports or statements.

Use of composite part scores. The total full-scale IQ score is usually the best estimate of a client's overall intellectual functioning for diagnostic purposes. However, there are instances in which, and individuals for whom, the total test score may not be the best representation of overall intellectual functioning. These situations occur when statistically significant and meaningful differences are observed between the part scores that comprise the full-scale IQ score. In such situations, appropriate part scores that have high *g*-loadings (e.g., WAIS-IV Verbal Comprehension; WJ III Fluid Reasoning and Comprehension-Knowledge) may be used when the validity of the full-scale IQ score is in doubt. There are certain additional, related considerations:

a. *Appropriate part or component scores as g-proxies.* The use of part scores to diagnose ID must be done cautiously and be based on sound scientific evidence and be consistent with accepted professional standards. The use of part scores may increase the potential for more accurate diagnosis for individual cases, but also raises the possibility of misuse via selective "cherry picking" of part scores to either support or refute an ID diagnosis. For example, as reflected in Table 7.1 and the associated discussion in this chapter, processing speed (*Gs*) scores are the weakest proxies for general intelligence. An argument that a person's low processing speed scores support a diagnosis of ID, in the context of higher *g*-loading scores (e.g., fluid and crystallized intelligence) that are above the ID range, is not supported by, and would be contrary to, the scientific evidence regarding *g*-ness of part or component scores. The presence of a high and significantly discrepant

WAIS-IV Processing Speed Index score that raises the WAIS-IV Full Scale IQ just above the ID cutoff score, when combined with WAIS-IV Perceptual Reasoning, Verbal Comprehension, and Working Memory Index scores within the ID range, would be consistent with the possibility of a diagnosis of ID despite the full-scale IQ score. Alternatively, a low WJ III General Intellectual Ability (GIA) score just within the ID range might mask a proper ID diagnosis if the individual had relative weaknesses (significant and meaningfully different) on the low-*g* WJ III Visual Processing and Processing Speed clusters, but strengths noticeably above the ID cutoff score on high-*g* Fluid Reasoning, Comprehension-Knowledge, and Long-Term Storage and Retrieval clusters. These two examples could be repeated with all the intelligence batteries listed in Table 7.1.

The part-score *g*-loadings presented in Table 7.1 provide initial guidance to assist assessment professionals in evaluating which part scores may be the best proxies for general intelligence within each respective intelligence battery. Examiners should seek out and use additional scientific evidence from each intelligence battery's technical manual and independent published research to support interpretations based on the *g*-ness of part scores in individual cases. Assessment professionals should be familiar with the *g*-loading scientific literature regarding those instruments when they use the pattern of part scores to support or refute a diagnosis of ID. An ID diagnosis based on part scores should be supported by the presentation of relevant research, as discussed earlier in the chapter, in written reports and statements.

b. *Significant and meaningful differences and patterns.* When part scores are used to formulate a diagnosis of ID, professionals must offer psychometric evidence that the full-scale IQ score is likely an invalid estimate of a person's general intelligence. Differential interpretation of part scores must only occur when the assessment professional provides evidence that the differences between part scores are *statistically significant* and *meaningfully different*. Part scores may be statistically significant, but the base rate in the population may not suggest that such a difference is meaningful (see Reschly et al., 2002, for detailed discussion). When the differences between part scores are not statistically significant or meaningful, or when the pattern of intellectual strengths and weakness does not display an internally consistent high and low *g*-loading part score pattern, then the full-scale IQ score should remain as the primary IQ score for estimating an individual's general intellectual functioning, as per the AAIDD definition.

Clinical judgment is often required and necessary in the interpretation of intelligence test results. Professional clinical judgment is often required and necessary when interpreting scores from intelligence batteries, particularly when an *Atkins* client has a file that contains multiple IQ test scores that span many years or when part scores are used (in place of the full-scale IQ score) as the basis of a diagnosis of ID.

Clinical judgment is a process based on solid scientific knowledge and is characterized as being "systematic (i.e., organized, sequential, and logical), formal (i.e., explicit and reasoned), and transparent (i.e., apparent and communicated clearly)" (Schalock & Luckasson, 2005, p.1). The misuse of clinical judgment in the interpretation of scores from intelligence test batteries should not be used as the basis for "gut instinct" or "seat-of-the-pants" impressions and conclusions of the assessment professional (Macvaugh & Cunningham, 2009), or justification for shortened evaluations, a means to convey stereotypes or prejudices, a substitute for insufficiently explored questions, or an excuse for incomplete testing and missing data (Schalock & Luckasson, 2005). Idiosyncratic methods and intuitive conclusions are not scientifically based and have unknown reliability and validity. If interpretations and opinions regarding an individual's level of general intelligence are based on novel or emerging research-based principles, the assessment professional must document the bases for these new interpretations as well as the limitations of these principles and methods.

Comparison of adaptive behavior and IQ scores. Intelligence test information must be interpreted within the context of relevant collateral information. Adaptive behavior is one major source of collateral information. Adaptive behavior total composite and intelligence full-scale IQ scores correlate at a moderate level. Thus, assessment professionals should not always expect adaptive behavior and IQ scores to be consistent. These two scores represent distinctly different measures of different domains of personal competence. Users can expect that, 68% of the time, an adaptive behavior composite score can range from as much as 15 points lower to 15 points higher than the measured full-scale IQ test score. At the 95% confidence level, the adaptive behavior composite scores may range from up to ±30 points different from any specific IQ score. When significant and meaningful adaptive behavior and intelligence test score differences are present for individuals, professionals must provide scientific and professionally accepted interpretation for differences that may include, but are not limited to:

a. adaptive behavior scales are measuring typical performance, while intelligence test batteries are measuring maximal performance;

b. adaptive behavior scales focus on everyday behavior, while intelligence tests emphasis mental thought processes;

c. adaptive behavior measures competencies that are more malleable and subject to change due to either positive or negative changes in a person's environment(s), while intelligence test batteries measure a more stable set of abilities; and

d. third-party informants typically provide the raw material of adaptive behavior in contrast to the individuals themselves providing direct information via their responses in a structured and standardized 1-1 testing situation.

See Chapters 11–13 for more detailed discussion of issues surrounding the administration and interpretation of adaptive behavior scales.

References

Ackerman, P. L., & Lohman, D. F. (2006). Individual differences in cognitive functions. In P. A. Alexander & P. Winne (Eds.), *Handbook of educational psychology*—(2nd ed.) (pp. 139–161). Mahwah, NJ: Erlbaum.

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

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

Brody, N. (2000). History of theories and measurements of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 16–33). New York, NY: Cambridge University Press.

Bruininks, R. H., Woodcock, R. W., Weatherman, R. F., & Hill, B. K. (1996). *SIB-R: Scales of Independent Behavior—Revised*. Chicago, IL: Riverside.

Burns, R. B. (1994). Surveying the cognitive terrain. *Educational Researcher, 23*(2), 35–37.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York, NY: Cambridge University Press.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38*, 592.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153–193.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research, 18*, 115–126.

Cudek, R., & MacCallum, R. C. (2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Erlbaum.

Daniel, M. H. (1997). Intelligence testing: Status and trends. *American Psychologist, 52*, 1038–1045.

Daniel, M. H. (2000). Interpretation of intelligence test scores. In R. Sternberg (Ed.), *Handbook of intelligence* (pp. 477–491). New York, NY: Cambridge University Press.

Drozdick, L. W., Wahlstrom, D., Zhu, J., & Weiss, L. G. (2012). The Wechsler Adult Intelligence Scale—(4th ed.) and the Wechsler Memory Scale—(4th ed.). In D. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 197–223). New York, NY: Guilford.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1979). Cognitive factors: Their identification and replication. *Multivariate Behavioral Research Monographs, 79*(2), 3–84.

Elliott, C. D. (2007). *Differential Ability Scales* (2nd ed.). San Antonio, TX: Pearson.

Erodi, L. A., Richard, D. C. S., & Hopwood, C. (2009). The importance of relying on the manual: Scoring error variance in the WISC-IV Vocabulary test. *Journal of Psychoeducational Assessment, 27*(5), 374–385. doi: 10.1177/0734282909332913

Flanagan, D. P., & Harrison, P. L. (Eds.). (2012). *Contemporary intellectual assessment* (3rd ed.). New York, NY: Guilford Press.

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

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice, 39*, 414–423. doi: 10.1037/0735-7028.39.4.414

French, J. W. (1951). *The description of aptitude and achievement tests in terms of rotated factors* (Psychometric Monographs No. 5). Chicago, IL: University of Chicago Press.

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Manual and kit of reference tests for cognitive factors*. Princeton, NJ: Educational Testing Service.

Gold, L. H., Anfang, S. A., Drukteinis, A. M., Metzner, J. L., Price, M., Wall, B. W., . . . & Zonana, H. V. (2008). AAPL practice guidelines for the forensic evaluation of psychiatric disability. *Journal of the American Academy of Psychiatry and the Law—Supplement, 36*(4), S3–S50. doi: 36/Supplement_4/S3

Guilford, J. P. (1967). *The nature of human intelligence.* New York, NY: McGraw-Hill.

Hakstian, A. R., & Cattell, R. B. (1974). The checking of primary ability structure on a basis of twenty primary abilities. *British Journal of Educational Psychology, 44,* 140–154.

Harrison, P. L. (1987). Research with adaptive behavior scales. *Journal of Special Education, 21,* 37–68.

Harrison, P. L., & Oaklan, T. (2003). Adaptive Behavior Assessment System manual (2nd ed.). Los Angeles: Western Psychological Services.

Hopwood, C. J., & Richard, D. C. S. (2005). WAIS-III scoring accuracy is a function of scale IC and complexity of examiner tasks. *Assessment, 12,* 445–454.

Horn, J. L. (1976). Human abilities: A review of research and theory in the early 1970s. *Annual Review of Psychology, 27,* 437–485.

Horn, J. L. (1989). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *WJ-R technical manual* (pp. 197–245). Chicago, IL: Riverside.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 53–91). New York, NY: Guilford.

Ittenbach, R., Spiegel, A., McGrew, K. S., & Bruininks, R. (1992). A confirmatory factor analysis of early childhood ability measures within a model of personal competence. *Journal of School Psychology, 30,* 307–323.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Jensen, A. R. (2004). Obituary—John Bissell Carroll. *Intelligence, 32,* 1–5. doi: 10.1016/j.intell.2003.10.001

Kaufman, A. S. (2009). *IQ testing 101.* New York, NY: Springer.

Kaufman, A. S., & Kaufman, N. L. (1993). Kaufman Adolescent and Adult Intelligence Test. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children* (2nd ed.). San Antonio, TX: Pearson.

Keith, T., Fehrmann, P., Harrison, P., & Pottebaum, S. (1987). The relation between adaptive behavior and intelligence. Testing alternative explanations. *Journal of School Psychology, 25,* 31–43.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47,* 635–650. doi: 10.1002/pits.20496

Kuentzel, J. G., Hetterscheidt, L. A., & Barnett, D. (2011). Testing intelligently includes double-checking Wechsler IQ scores. *Journal of Psychoeducational Assessment, 29,* 39–46. doi: 10.1177/0734282910362048

Lichtenberger, E. O., & Kaufman, A. S. (2009). *Essentials of WAIS-IV assessment.* Hoboken, NJ: Wiley.

Macvaugh, G. S., & Cunningham, M. D. (2009). *Atkins v. Virginia*: Implications and recommendations for forensic practice. *Journal of Psychiatry and Law, 37,* 131–187.

Meyers, C. E., Nihira, K., & Zetlin, A. (1979). The measurement of adaptive behavior. In N R. Ellis (Ed.), Handbook *of mental deficiency: Psychological theory and research* (2nd ed., pp. 215–253). Hillsdale, NJ: Erlbaum.

McGrew, K. S. (1994). *Clinical interpretation of the Woodcock-Johnson Tests of Cognitive Ability Revised*. Boston, MA: Allyn and Bacon.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–179). New York, NY: Guilford.

McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–181). New York, NY: Guilford Press.

McGrew, K. (2009). Editorial: CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1–10. doi:10.1016/j.intell.2008.08.004

McGrew, K. (2010). *Applied psychometrics 101 brief #6: Understanding Wechsler IQ score differences—the CHC evolution of the Wechsler FS IQ score*. Retrieved from http://www.iqscorner.com/2010/02/ap101-brief-6-understanding-wechsler-iq.html

McGrew, K. S. (2012). IAP Applied Psychometrics 101 Brief Report # 11: What is the typical IQ and adaptive behavior correlation? *Intelligent Insights on Intelligence Theories and Tests (aka, IQ's Corner)*. Retrieved from http://www.iqscorner.com/search/label/adaptive%20behavior

McGrew, K., & Bruininks, R. (1990). Defining adaptive and maladaptive behavior within a model of personal competence. *School Psychology Review, 19*, 53–73.

McGrew, K., & Flanagan, D. (1998). *The Intelligence Test Desk Reference (ITDR). Gf-Gc cross-battery assessment*. Boston, MA: Allyn & Bacon.

Mental Measurement Yearbook. Retrieved from http://buros.org/how-cite-reviews-buros-institutes-test-reviews-online

Naglieri, J. A., Das, J. P., & Goldstein, S. (2012). Planning, attention, simultaneous, successive: A cognitive-processing-based theory of intelligence. In D. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 178–194). New York, NY: Guilford.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., ... Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Phelps, L., McGrew, K. S., Knopik, S. N., & Ford, L. (2005). The general (*g*), broad, and narrow CHC stratum characteristics of the WJ III and WISC-III tests: A confirmatory cross-battery investigation. *School Psychology Quarterly, 20*, 66–88.

Ramos, E., Alfonso, V. C., & Schermerhorn, S. M. (2009). Graduate students' administration and scoring errors on the Woodcock-Johnson III Tests of Cognitive Abilities. *Psychology in the Schools, 46*, 650–657. doi: 10.1002/pits.20405

Raven, J. C. (1976). *Standard progressive matrices*. Oxford, England: Oxford Psychologists Press.

Reschly, D., Myers, T., & Hartel, C. (2002). *Mental retardation: Determining eligibility for social security benefits*. Washington, DC: National Academy Press.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scale* (5th ed*)*. Austin, TX: Pro-Ed.

Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring reliability on the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III). *Assessment, 10*, 151–159. doi: 10.1177/1073191103010002006

Sattler, J. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Jerome M. Sattler, Publisher, Inc.

Schalock, R. L., Buntinx, W. H. E., Borthwick-Duffy, S., Bradley, V., Craig, E. M., Coulter, D. L., Gomez, S. C., Lachapelle, Y., Luckasson, R. A., Reeve, A., Shogren, K. A., Snell, M. E., Spreat, S., Tassé, M. J., Thompson, J. R., Verdugo, M. A., Wehmeyer, M. L., & Yeager, M. H. (2010). *Intellectual disability: Definition, classification, and system of supports* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schalock, R. L. & Luckasson, R. (2005). *Clinical judgment*. Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schalock, R. L., Luckasson, R., Bradley, V., Buntinx, W., Lachapelle, Y, Shogren, K. A., Snell, M. E., Thompson, J. R., Tassé, M., Verdugo-Alonso, M. A., & Wehmeyer, M. L. (2012). *User's guide to mental retardation: Definition, classification, and systems of supports*. Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schneider, W. J., & McGrew, K. (2012). The Cattell-Horn-Carroll model of intelligence. In D. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*(3rd ed.; pp. 99–144). New York, NY: Guilford.

Singer, J. K., Licthenberger, E. O., Kaufman, J. C., Kaufman, A. S., & Kaufman, N. L. (2012). The Kaufman Assessment Battery for Children (2nd ed.) and the Kaufman Test of Educational Achievement (2nd e.d,) In D. Flanagan, & Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.; pp. 178–194). New York, NY: Guilford.

Snow, R. E. (1998). Abilities and aptitudes and achievements in learning situations. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 93–112). Mahwah, NJ: Erlbaum.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland—II: Vineland Adaptive Behavior Scales* (2nd ed.). San Antonio, TX: Pearson.

Spearman, C. E. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Spearman, C. E. (1927). *The abilities of man: Their nature and measurement.* London, England: Macmillan.

Sternberg, R. J., & Kaufman, J. C. (1998). Human abilities. *Annual Review of Psychology, 49*, 479–502.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1*.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometrics Monographs, 2*.

VandenBos, G. (2007). *APA dictionary of psychology*. Washington, DC: American Psychological Association.

Wasserman, J. D. (2012). A history of intelligence assessment: The unfinished tapestry. In D. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.; pp. 3–55). New York, NY: Guilford.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: Pearson.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children* (3rd ed.). *New York: Psychological Corporation.*

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

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

Whitaker, S. (2008). The stability of IQ in people with low intellectual ability: An analysis of the literature. *Intellectual and Developmental Disabilities, 46*(2), 120–128. doi: 10.1352/0047-6765(2008)46[120:TSOIIP]2.0.CO;2

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson III Normative Update.* Rolling Meadows, IL. Riverside.