CHAPTER

4

Psychometric Versus Actuarial Interpretation of Intelligence and Related Aptitude Batteries

Gary L. Canivez

Abstract

Interpretation of intelligence tests involves making various inferences about the individual based on their performance. Because there are many different scores within intelligence tests reflecting different levels (Full Scale, factors, subtests) and there are many different comparisons provided in test manuals and the extant literature, there is a multitude of possible inferences. This chapter is concerned with reviewing the various available scores and comparisons with suggested interpretations and a review of the empirical investigations of their psychometric fitness (reliability, validity, utility). Differentiation of psychometric interpretation versus actuarial interpretation methods is presented, as well as a review of research related to each. Most intelligence test interpretation methods are considered psychometric in nature, and most lack sufficient reliability, validity, or utility for individual clinical use; improvements in the clinical assessment of intelligence may result from greater development and use of actuarial approaches.

Key Words: intelligence test interpretation, reliability, validity, utility, actuarial decision-making, clinical decision-making

Introduction

Interpretation of intelligence tests involves drawing inferences about an individual based on scores obtained on a particular test. Because contemporary intelligence tests provide many different types of scores, there is a variety of inferences that could be made about any individual. Furthermore, various intelligence tests are constructed to reflect different theories of intelligence or cognitive abilities, and interpretations may also be based on the particular theory upon which the test is based. While a test may be created to reflect a particular theory, clinicians may also apply alternate or competing theories in the interpretation of test scores. Legitimate inferences about an individual from various intelligence test scores or procedures, however, must *each* be supported by reliability, validity, and utility research on the various scores, comparisons, and their uses.

Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education [AERA, APA, NCME], 1999) provides numerous guidelines for considering the reliability and validity of test scores that should be applied to intelligence test scores. Such guidelines apply to test authors and publishers, but ultimately it is the test user who must decide which test scores, comparisons, and procedures possess sufficient evidence of reliability and validity to report and interpret. Test scores that do not possess adequate reliability, validity, and utility will lead the test user to make inaccurate and inappropriate inferences about the individual when interpreting those test scores and comparisons. Such inaccurate and inappropriate inferences may well lead to recommendations for classification,

diagnosis, or treatment that may also be wrong. Weiner (1989) cogently noted that in order to practice in an ethical manner, psychologists must "(a) know what their tests can do and (b) act accordingly" (p. 829). Numerous ethical standards also concern the use of tests and measurement procedures in clinical practice (APA, 2010; National Association of School Psychologists [NASP], 2010).

Interpretation of intelligence tests may involve description of the individual's performance, prediction of the individual's performance in related areas such as classroom learning or performance on academic achievement tests, classification or diagnosis of the individual, and informing or recommending treatments. Each of these "interpretations" requires empirical support for appropriate use. With respect to individual test scores, it has been argued that if a score is to be used for individual decisions or clinical decision-making, reliability indices should meet or exceed .85 (Hills, 1981) or .90 (Aiken, 2000; Guilford & Fruchter, 1978; Nunnally & Bernstein, 1994; Ponterotto & Ruckdeschel, 2007; Salvia & Ysseldyke, 1988, 2001). Specific inferences about the individual are tied more to the various estimates of the validity of test scores and their diagnostic utility; however, scores that lack sufficient reliability cannot be valid or of diagnostic utility. It is within this framework that psychometric and actuarial intelligence test interpretations are examined.

History of IQ Test Interpretations

Kamphaus, Winsor, Rowe, and Kim (2005) provided a useful description of the history of intelligence test interpretation. The earliest application of intelligence test interpretation was associated with a classification of the individual's test score (and thus the individual) according to descriptive terms based on the overall test score and prediction of school functioning. The earliest classification systems contained descriptive terms (viz., "idiots," "imbeciles," "morons"; Levine & Marks, 1928) that are considered pejorative by today's standards. Each intelligence test presently published also contains descriptive classifications for test score ranges, and these are frequently the first "interpretation" made. Kamphaus et al. also noted that present-day descriptive terms typically reflect some aspect of score deviation from the average range, but each test contains somewhat different descriptors and may include different score ranges. While the earliest intelligence tests provided one overall test score (i.e., IQ), present-day intelligence tests contain numerous scores reflecting different levels of the test. Such scores include an overall, omnibus Full Scale score; several factor-based scores or indexes; subtest scores; ipsative or deviation scores derived from comparing subtest or factor scores to the individual's overall mean score; and theoretically or logically combined subtest composite scores.

Kamphaus et al. (2005, p. 26) referred to a "second wave" of test interpretation proposed by Rapaport, Gil, and Schafer (1945–1946) that ushered in an approach to intelligence test interpretation still in use by many today. The approach advocated by Rapaport et al. was that of going beyond the overall, omnibus Full Scale IQ and examining the shape of the subtest profile to provide a description of subtest highs and lows for the individual. These strengths and weaknesses were presumed to reflect some aspect of psychopathology as well as consideration for intervention. Wechsler's examination (1944) of differences between verbal and performance scales as well as subtest profile shape and deviations was also included in this second wave and reflected Wechsler's clinical approach to test interpretation. This is in contrast to Wechsler's note that subtests are merely different practical estimates for measuring general intelligence.

Criticism of these early interpretation methods on empirically based psychometric grounds led to what Kamphaus et al. (2005) referred to as the "third wave," wherein application of psychological measurement methods was used to evaluate various intelligence test scores and interpretation methods. Cohen's investigation (1959) of the factor structure of the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) was one of the first of its kind and provided an empirical means for identifying subtest association to factors or dimensions underlying the WISC. While Wechsler assigned WISC subtests to the verbal or performance scale based on subtest content, Cohen's factor analysis empirically assigned subtests to factors based on their shared variance, and resulted in the three-factor structure (Verbal Comprehension, Perceptual Organization, and Freedom from Distractability) of the WISC. Cohen was critical of WISC subtest scores based on the high levels of shared variance and low subtest specificity (non-error variance unique to the subtest). Subsequently, this three-factor structure was frequently replicated with other samples and Cohen's WISC factor names were retained in both the second (WISC-R; Wechsler, 1974) and third editions of the WISC (WISC-III; Wechsler, 1991). Kaufman (1979) provided a means for calculating

scores for the three factors to provide more factorially pure scores than the VIQ and PIQ.

Kaufman's significant influence on intelligence test interpretation is reflected in the hierarchical and sequential process of interpretation and subtest analyses (Kaufman, 1979, 1994; Kaufman & Lichtenberger, 2000, 2006) frequently taught in graduate programs and used by practitioners (Alfonso, Oakland, LaRocca, & Spanakos, 2000; Groth-Marnat, 1997; Kaufman, 1994; Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). Levels of test interpretation were ordered from the most reliable and valid scores (Full Scale IQ and composite scores) to the least reliable and valid scores (single subtest scores). Kaufman's approach continued the clinical interpretation method of unique profile shape (subtest strengths and weaknesses) much like that of Rapaport et al. (1945–1946), but his profile interpretation method of comparing an individual's subtest scores to their own overall mean performance was through the identification of statistically significant strengths and weaknesses and consideration of base rates. This ipsative approach examines intraindividual differences and is an *ideographic* interpretation approach, in contrast to the *nomothetic* interpretation approach (normative comparisons) provided by the standard scores of intelligence tests. Sattler (1982, 1988, 1992, 2001, 2008) and Sattler and Ryan (2009) also provided for similar intelligence test interpretations based on a sequential order from the global score to subtest comparisons.

Reliability and validity research highlighting major problems with ipsative comparisons and profile interpretations (discussed in detail later in this chapter) led to what Kamphaus et al. (2005) referred to as the "fourth wave" where intelligence theory was applied to intelligence test construction and interpretations. The earliest intelligence tests were seemingly constructed from a pragmatic perspective and some would say atheoretical. Thorndike (1990, p. 226) noted Alfred Binet was "theoretically agnostic" with regard to the Binet-Simon scale. Zachary (1990) noted David Wechsler's minimal explication of theory in constructing the original Wechsler-Bellevue (WB) scales in 1939. However, as several have pointed out, when constructing and modifying his tests, Wechsler used Spearman's (1904, 1927) theory of general intelligence (*g*) and was also influenced by other contemporary intelligence theories (Saklofske, 2008; Tulsky et al., 2003; Zhu & Weiss, 2005). Wechsler's definition of *intelligence* (i.e., "global capacity"; Wechsler, 1939, p. 229) also reflected Spearman's *g*.

Contemporary intelligence tests are now greatly influenced by theories of intelligence or models of intelligence measurement in their construction and interpretation. Presently, one of the most influential models of intellectual measurement is that of Carroll (1993, 1995, 1997a, 2003), which proposes that intelligence tests measure various intellectual abilities that are hierarchically ordered. Carroll's (1993, 2003) three-stratum theory of cognitive abilities proposes some 50–60 narrow abilities (Stratum I) at the bottom (subtests), 8–10 broad-ability factors (Stratum II) in the middle (first-order factors), and the general ("g") ability factor (Stratum III) at the top (second-order factor) and reflected by the overall FSIQ or global IQ. Many contemporary intelligence tests are constructed to reflect Carroll's model of intelligence measurement either explicitly or implicitly. The American Psychological Association task force study of intelligence (Neisser et al., 1996) noted that the hierarchical nature of intelligence measurement was the most widely accepted view, and this still appears to be true.

Another theoretical perspective of intelligence measurement closely related to, and preceding that of Carrol, is the G*f*–G*c* theory of Cattell and Horn (Cattell, 1943; Horn, 1988, 1991; Horn & Cattell, 1966; Horn & Noll, 1997). Cattell and Horn's G*f*–G*c* theory describes aspects of reasoning abilities that allow the individual to solve novel problems (*fluid intelligence* [G*f*]) and abilities acquired through the individual's exposure to aspects of their culture such as language and educational experiences (*crystallized intelligence* [G*c*]). Extension of G*f*–G*c* theory by Horn (1991) and Horn and Noll (1997) is similar to that of Carroll with some 8–9 or more broad dimensions but does not include higher-order *g,* arguing there was insufficient construct validity evidence for singular *g*.

Cattell-Horn-Carroll (CHC) theory is an approach wherein researchers melded the work of Cattell and Horn with that of Carroll (Evans, Floyd, McGrew, & Leforgee, 2001; Flanagan, 2000; McGrew, 2005), but this is an odd combination, given that Carroll provided evidence for higher-order *g* while Horn argued that singular *g* did not exist and was a statistical artifact. Carroll's model, Cattell and Horn's model, or the combined CHC model are often cited as theoretical foundations or influences in present versions of intelligence tests such as the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV: Wechsler, 2003), Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV: Wechsler, 2008a), Stanford-Binet

Intelligence Scales–Fifth Edition (SB-5: Roid, 2003a), Differential Ability Scales–Second Edition (DAS-II: Elliott, 2007a), Wide Range Intelligence Test (WRIT: Glutting, Adams, & Sheslow, 2000a), Reynolds Intellectual Assessment Scales (RIAS: Reynolds & Kamphaus, 2003), Kaufman Adolescent and Adult Intelligence Test (KAIT: Kaufman & Kaufman, 1993), and Kaufman Assessment Battery for Children–Second Edition (KABC-II: Kaufman & Kaufman, 2004a).

Luria's neuropsychological theory was the foundation for the development of the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983) and later, the development of the Planning, Attention, Simultaneous, Successive (PASS) theory (Das, Naglieri, & Kirby, 1994) which is the foundation for the Cognitive Assessment System (CAS: Naglieri & Das, 1997a). While the KABC-II is linked to and can be interpreted according to CHC theory, it is also linked to and can be interpreted according to Luria's theory through its dual theoretical foundation (Kaufman & Kaufman, 2004a). CAS interpretation follows the PASS theory and is related to Luria's three functional units of the brain (Naglieri, 1997; Naglieri & Das, 1990, 1997b).

Kamphaus et al. (2005) noted that, in practice, clinicians frequently interpret not only the overall Full Scale/Composite intelligence test score, which is an estimate of Spearman's *g* (Spearman 1904, 1927) and the apex of Carroll's hierarchy (1993), but also interpret factor-based scores *and* ability profiles produced from subtest strengths and weaknesses. These subtest profile-interpretation systems (Flanagan & Kaufman, 2004; Kaufman, 1994; Kaufman & Lichtenberger, 2000; Sattler, 2001, 2008; Sattler & Ryan, 2009) are exceedingly popular in psychology training and clinical practice (Alfonso et al., 2000; Groth-Marnat, 1997; Kaufman, 1994; Pfeiffer et al., 2000).

Despite their popularity, clinicians *must* confront the issue of whether there is sufficient empirical support for *all* these interpretation methods and the resulting inferences. While many test publishers, test authors, workshop presenters, and textbook authors present all the above methods for interpreting intelligence tests and argue for their utility, it is the test user who must decide which methods have adequate reliability, validity, and utility for clinical application and making inferences and decisions about the individual they assess. By interpreting all such scores, the clinician is adding (or replacing) interpretations beyond the overall Full Scale score and thus, knowledge of such score interpretation improvement *beyond* the overall IQ score is imperative (Brody, 1985). Furthermore, interpretation of scores at every level of the test ignores the fact that scores at the subtest and factor level are themselves correlated and *not* independent, representing mixtures of *g* and subtest variance (Carroll, 1993). These and other critically important interpretation issues are further explicated throughout this chapter.

Psychometric Interpretation Methods

Descriptions of intelligence test interpretation methods from stage one through stage four (Kamphaus et al., 2005) are ostensibly *psychometric* interpretation methods. Interpretations based on psychometric grounds are interpretations based on various scores provided by the intelligence test as well as derived scores or comparisons of different scores. These scores can be used in descriptive (level of performance), predictive (estimate of performance on related variable), and classificatory (assignment to diagnostic group) ways. Clinicians use test scores and analysis information as well as test session observations, background information, interview information, etc., and make a clinical decision, judgement, or inference regarding the individual. Such decisions might be considered *clinical* decisions rather than *actuarial* (statistical) decisions because it is the clinician's *judgement* of meaning of the score(s) that guides the interpretation (decision) rather than strict adherence to a statistically based (formula) interpretation (decision) (Meehl, 1954, 1957; Meehl & Rosen, 1955).

Intelligence test scores are (or can be) associated with a hierarchical model similar to Carroll's (1993, 2003) Stratum III (omnibus, Full Scale score), Stratum II (factor-based scores), or Stratum I (subtest scores). These scores are standardized scores and reflect comparison to the normative group appropriate for the individual. Interpretations from the standardized scores are nomothetic and allow for understanding how the individual performed relative to others their age in the population. Other derived scores such as ipsative factor score comparison, pairwise factor score comparisons, ipsative subtest comparisons, and pairwise subtest comparisons are rooted in the clinical intelligence test interpretation approaches articulated by Rapaport et al. (1945–1946), Kaufman (1979, 1994), Kaufman and Lichtenberger (2000, 2006), Sattler (1982, 1988, 1992, 2001, 2008), and Sattler and Ryan (2009) and are ideographic. These are comparisons of the individual to himself or herself. Another

means by which clinicians may interpret scores from intelligence tests is the combination of various subtest scores into composite scores based on some logical or theoretical connection (Kaufman, 1994; Kaufman and Lichtenberger, 2000, 2006; Sattler, 2001, 2008; Sattler & Ryan, 2009).

Finally, some have argued that no single intelligence test adequately measures all of Carroll's (1993, 2003) Stratum II dimensions, or Horn and Noll's (1997) G*f*–G*c* factors, and in order to better assess these broad intelligence dimensions, subtests from different intelligence tests should be combined in what is referred to as "cross-battery assessment" (Flanagan & McGrew, 1997; Flanagan & Ortiz, 2001; Flanagan, Ortiz, & Alfonso, 2008; McGrew & Flanagan, 1998). Cross-battery assessment is rooted in the aforementioned CHC theory. Such cross-battery assessment is proposed to better account for an individual's varied broad abilities (Stratum II) in cognitive assessment; which in turn is used to understand their educational difficulties, provide better differential diagnosis, and guide interventions. While the cross-battery assessment approach attempts to improve assessment of cognitive abilities and has intuitive appeal, there are numerous substantial psychometric problems cogently pointed out by Glutting, Watkins, and Youngstrom (2003), which have yet to be adequately empirically addressed. Other issues noted later in this chapter also have implications for such interpretations. Thus, individual clinical use of such cross-battery interpretation methods is not recommended.

Examination of psychometric interpretation methods proceeds in order of the hierarchy and structure of intelligence tests (Carroll, 1993, 1995; Neisser et al., 1996). This order also parallels the psychometric interpretation procedures recommended by Kaufman (Flannagan & Kaufman, 2004; Kaufman, 1994; Kaufman & Lichtenberger, 2002, 2006) and Sattler (Sattler, 2008; Sattler & Ryan, 2009) as well as contemporary intelligence test authors through their test technical and interpretation manuals.

Global IQ Score Interpretation/Stratum III

The first level of intelligence test interpretation involves the reporting and description of the overall, omnibus, global, or Full Scale score; which is measured or estimated by two or more subtests within the scale. Full Scale scores represent an estimate of *g* or Spearman's general intelligence factor (Spearman, 1904, 1927), which is the apex of Carroll's (1993, 1995) hierarchical model. Substantial evidence for interpretation of global Full Scale scores (*g* estimates) exists (Gottfredson, 2002, 2008; Jensen, 1998; Kubiszyn et al., 2000; Lubinski, 2000; Lubinski & Humphreys, 1997; Neisser et al., 1996).

Interpretation of Full Scale scores typically begins with a presentation and description of the standard score, percentile rank, and confidence interval (obtained score or estimated true score) for the standard score to account for measurement error, consistent with *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). Classification of the Full Scale score within a descriptive category (or range of descriptive categories around the confidence interval) reflecting some deviation from average is also frequently made (i.e., average, above average, significantly below average). These normative descriptors are intended to provide an illustration of how the individual performed relative to others their age. This is necessary given the interval level measurement such standardized scores represent.

Another aspect of interpretation of the Full Scale score is an inference as to what to expect from the individual in regard to their acquisition of academic skills (i.e., academic achievement). The predictive validity of intelligence test Full Scale scores is well documented (Bracken & Walker, 1997; Brody, 2002; Brown, Reynolds, & Whitaker, 1999; Flanagan, Andrews & Genshaft, 1997; Gottfredson, 2008, Konold & Canivez, 2010; Naglieri & Bornstein, 2003) and intelligence is a construct that precedes and influences the development of academic achievement because as Jensen (1998) noted "school learning itself is *g*-demanding" (p. 279). Watkins, Lei, and Canivez (2007) also demonstrated that verbal and perceptual measures of intelligence (WISC-III) predicted future academic achievement in reading and mathematics but reading and mathematics *did not* predict future measured intelligence, thus supporting Jensen's position of the temporal precedence and influence of intelligence on academic achievement.

Full Scale scores are also used in making classification or diagnostic decisions regarding psychopathology, disability, and giftedness. Classification or diagnosis of mental retardation (MR)/intellectual disability (ID) requires by definition "significantly subaverage general intellectual functioning" (Public Law 108–446 [IDEIA]; U. S. Department of Education, 2006, p. 46756) or "significantly subaverage intellectual functioning" (APA, 2000, p. 49). It is generally agreed that a score 2 *SD* below

the population mean would satisfy this criterion. Historically, the operational definition for classifying specific learning disability (SLD), according to IDEA prior to IDEIA 2004, involved first a severe discrepancy between predicted academic achievement given a certain level of intelligence and actual academic achievement. This criterion was concerned with the *unexpectedly* low achievement associated with the concept of learning disability (Reynolds, 1984). While federal law no longer mandates the use of the severe discrepancy criterion, states may allow local school districts to continue using this criterion in SLD classification (IDEIA 2004) and predicted (expected) achievement is based on a person's general intellectual ability. DSM-IV-TR (APA, 2000) specifies individually measured achievement from a standardized test that "is substantially below that expected given the person's chronological age, measured intelligence, and age-appropriate education" (p. 53, p. 54, p. 56) for classification of its SLD type disorders (viz., Reading Disorder, Mathematics Disorder, Disorder of Written Expression, respectively). Assessment for intellectual giftedness also has implications for intelligence testing and it is generally thought that a Full Scale intelligence test score 2*SD* above the population mean would satisfy the criterion of significantly above average intelligence.

GLOBAL SCORE PSYCHOMETRIC SUPPORT

Psychometric support for Full Scale IQ, or related omnibus scores, is strong and includes the highest internal consistency estimates, short-term temporal stability, long-term temporal stability, and predictive validity coefficients (Bracken & McCallum, 1998b; Canivez & Watkins, 1998; Elliott, 2007b; Glutting, Adams, & Sheslow, 2000b; Kaufman & Kaufman, 1983, 1993, 2004a, 2004b; Naglieri & Das, 1997b; The Psychological Corporation, 1999; Reynolds & Kamphaus, 2003; Roid, 2003b; Wechsler, 2002a, 2002b, 2003, 2008b; Wechsler & Naglieri, 2006; Woodcock, McGrew, & Mather, 2001). In studies of long-term temporal stability of WISC-III IQ scores, the Full Scale IQ possessed the highest stability coefficient of all scores available and this also held across a variety of demographic variables (sex, age, race/ethnicity, disability) (Canivez & Watkins, 1998, 1999, 2001). Similar results were also reported by Krohn and Lamp (1999) for the K-ABC and Stanford-Binet Intelligence Scale: Fourth Edition (SB:FE: Thorndike, Hagen, & Sattler, 1986). These are expected findings according to true score theory, and reflect the power of aggregate scores that contain less error variance. Assessing temporal stability is important as it addresses one of the major sources of error variance in intelligence tests not addressed by the internal consistency estimate (Hanna, Bradley, & Holen, 1981). Hanna et al. also noted the importance of assessing intelligence test measurement error related to scoring and administration errors.

IQ tests have a rich history of accounting for meaningful levels of academic achievement variance (Brody, 2002; Naglieri & Bornstein, 2003), with average IQ-achievement correlations near .55 across age groups (Neisser et al., 1996; Brody, 2002). Thorndike (1986) also noted approximately 85% to 90% of predictable criterion variable variance is accounted for by a single general score. Among co-normed intelligence and achievement tests it is quite common to observe concurrent IQ-achievement correlations near .70. It is often said that the most important application of intelligence tests is their ability to forecast student achievement (Brown, Reynolds, & Whitaker, 1999; Weiss & Prifitera, 1995) and the prediction of school performance with intelligence tests has been a primary use since the creation of the first Binet-Simon Scale of Intelligence (Binet & Simon, 1905).

Broad Factor/Verbal-Nonverbal Score Interpretation/Stratum II

VERBAL VS. NONVERBAL ASSESSMENT

Recognition of literacy issues affecting standardized intelligence testing was noted in the early 1900s and influenced the creation and use of the Army Alpha (verbal) and Army Beta (nonverbal) tests (Thorndike, 1997), with Yoakum and Yerkes (1920) noting the use of Army Beta for recruits who failed Army Alpha to prevent "injustice by reason of relative unfamiliarity with English" (p. 19). Wechsler (1939) also recognized the need to assess intellectual abilities through both verbal *and* nonverbal (performance) means. The selection and aggregation of subtests into composite scores were originally based on subtest task requirements, and, Wechsler noted, "the subtests are different measures of intelligence, not measures of different kinds of intelligence" (1958, p. 64). For Wechsler, verbal and performance measures were not different types of intelligence, but rather different ways to measure it. Verbal and performance measures relate to Carroll's (1993) Stratum II dimensions, which he referred to as "flavors" of *g*. Authors of contemporary *nonverbal* intelligence tests note that it is the method of assessment that is nonverbal rather than the cognitive process involved in solving the tasks (Bracken &

McCallum, 1998a, 1998b; Naglieri, 2003a, 2003b; Wechsler & Naglieri, 2006). Naglieri wrote, "the term *nonverbal* refers to the content of the test, not a type of ability" (2003a, p. 2). Thus, the content or demands of subtests and composites may differ but they still measure general intelligence. Spearman referred to this as the "indifference of the indicator" (1927, p. 197).

For individuals who have hearing impairments or are deaf, have receptive or expressive language deficits, or are from ethnic minority groups with limited English proficiency; nonverbal assessment of intellectual abilities is particularly useful (Bracken & McCallum, 1998b; Naglieri, 2003a). This is primarily because the verbal (English) methods would likely underestimate the intellectual abilities of such individuals. But there are other clinical groups for whom differences between verbal and nonverbal (performance) estimates have been reported or hypothesized. These include individuals with traumatic brain injury, bilingualism, autistic disorder, Asperger's disorder, and delinquents or psychopaths. Most intelligence test technical manuals (i.e., WPPSI-III, WISC-IV, WAIS-IV, KAIT, UNIT, KABC-II, DAS-II) typically provide clinical group and matched normal group comparisons in test performance that sometimes illustrate similar verbal–nonverbal differences but frequently contain small samples that are intended to be only preliminary investigations.

Verbal–Nonverbal (VIQ-PIQ/VCI-PRI) Comparisons

Kaufman and Lichtenberger (2006) devoted two chapters to VIQ and PIQ differences as they related to neuropsychology (brain functioning or injury) and clinical research and use. Many of the studies were of adults and with WB or WAIS data, but some studies were with children and adolescents and WISC data. Their review of more than 50 studies and approximately 2,700 patients who had unilateral brain damage generally supported the PIQ > VIQ for left-hemisphere lesions and VIQ > PIQ for right-hemisphere lesions on the WB, WAIS, and WAIS-R, but there was greater consistency for the right-hemisphere lesion groups. Variables that appeared to have possible moderating effects included age, sex, race or ethnicity, and educational attainment. Group difference studies are not sufficient for individual diagnostic use of such signs, and Kaufman and Lichtenberger noted calls by Matarazzo and colleagues (Bornstein & Matarazzo, 1982, 1984; Matarazzo, 1972; Matarazzo & Herman, 1984) for due caution for such individual use of VIQ-PIQ differences by attending to base rates in the population. Furthermore, it is one thing to assess an individual with known brain injury and lesion, observe VIQ-PIQ differences of a large magnitude, and infer the likely cause for the VIQ-PIQ difference to the brain damage; but in the absence of brain injury or lesion, to infer such from VIQ-PIQ differences is a much riskier proposition.

Such inferential problems from test scores or other "signs" in psychology have been pointed out at least as far back as 1955 (Meehl & Rosen, 1955) and are also articulated by others (Lilienfeld, Wood, & Garb, 2000; McFall 2005; Swets, Dawes, & Monahan, 2000; Watkins, 2009; Watkins, Glutting, & Youngstrom, 2005) as it relates to the Reverend Thomas Bayes' (1702–1761) theorem of conditional probabilities and base rates. Leonard and Hsu (1999) and Nickerson (2004) provide excellent descriptions of Bayes' theorem (Bayes, 1763) and its implications and applications. One feature of diagnostic tests often highlighted is *sensitivity,* which indicates the probability of obtaining a positive test result, given that the person has the target disorder. However, in diagnostic use of a test, a clinician is much more interested in *positive predictive power,* or the probability of a person's having the target disorder, given a positive test result. A similar contrast is that of *specificity,* which indicates the probability of obtaining a negative test result, given that the person *does not* have the target disorder; and *negative predictive power*, the more important indicator of the probability of a person's not having the target disorder, given a negative test result. With respect to low base rates, it is difficult for tests to improve accuracy in individual cases (Lilienfeld et al., 2006; McFall, 2005; Meehl, 2001; Meehl & Rosen, 1955).

In the case of VIQ-PIQ differences and inferences regarding brain injury and function, inverse probabilities suggest that there may well be a much greater proportion of individuals with brain injury and lesions who show VIQ-PIQ differences (sensitivity) than individuals with VIQ-PIQ differences who also have brain injury and lesions (positive predictive power). Neuropsychologists are also probably more likely to see patients referred for evaluations who have brain damage and observed VIQ-PIQ (VCI-PRI) discrepancies; and while those who have brain damage are more likely to show VIQ-PIQ (VCI-PRI) discrepancies, neuropsychologists may overestimate the value of these VIQ-PIQ (VCI-PRI) differences because they are not likely to

see those with VIQ-PIQ (VCI-PRI) discrepancies who do not have brain damage.

Review and summary of research provided by Kaufman and Lichtenberger (2006) regarding PIQ > VIQ for delinquents or "psychopaths" indicated that some distinct group difference studies found such differences as suggested by Wechsler (1944, 1958), but results overall were reportedly mixed. Kaufman and Lichtenberger noted that use of this PIQ > VIQ "sign" as recommended by Wechsler should not be used for individual diagnosis due to a lack of empirical support. Inconsistency of PIQ > VIQ findings for individuals with autistic disorder was also noted, in addition to small effect sizes, and thus determined to be of no diagnostic clinical use (Kaufman & Lichtenberger). More recently, a large Swedish study of individuals with Asperger's disorder, autism, or pervasive developmental disorder–not otherwise specified (PDD-NOS), based on DSM-IV criteria (American Psychiatric Association, 1994), found profile level (overall IQ) differentiated Asperger's from autism and PDD-NOS, but scatter and shape of profiles were small (Zander & Dahlgren, 2010). Within the autism group, a mean VCI-POI difference of 9 points (SD = 20.5) in favor of POI was observed and represented a medium effect size, but Zander and Dahlgren also noted individual profiles were too variable for individual diagnostic use of the Swedish version of the WISC-III (Wechsler, 1999) in differentiating among PDD diagnoses.

Cronbach (1990), however, noted problems with interpreting difference scores due to their low reliability. VIQ-PIQ difference scores, for example, have been shown to have poor temporal stability (too low for individual clinical use) and thus likely of questionable utility (Canivez & Watkins, 1998, 1999, 2001; Cassidy, 1997). The inference from significant VIQ-PIQ differences is that the individual has stronger cognitive skills in one area than the other, as well as giving rise to speculation as to the implications of the difference. However, if the difference score is not sufficiently reliable, it cannot be valid or of diagnostic value. Also, what such an analysis and inference ignores is the fact that VIQ and PIQ scores are not independent and such inferences from them are obscured by shared variance.

While Canivez, Neitzel, and Martin (2005) did not examine VIQ-PIQ differences in their study on relationships between the WISC-III, KBIT, Adjustment Scales for Children and Adolescents (ASCA: McDermott, Marston, & Stott, 1993), and academic achievement; with a sample (N = 207) of various students (non-disabled, learning disability, mental retardation/intellectual disability, and emotional disability); additional VIQ-PIQ analyses of this dataset were conducted for this chapter. Correlations between VIQ-PIQ discrepancies and all ASCA syndromes (core and supplementary) and global adjustment scales ranged from –.004 to .056 (p >.05) and ranged from –.003 to .099 (p >.05) with measures of academic achievement. Furthermore, there were no significant differences in VIQ-PIQ discrepancies between the four diagnostic groups. Glutting, Youngstrom, Oakland, and Watkins (1996) also examined relations between WISC-III scores and ASCA (and other measures) and found low WISC-III IQ and index score correlations ranging from –.27 to .18 (M_r = –.04) across all ASCA syndromes and global adjustment scales.

While some research may indicate distinct group differences with respect to VIQ > PIQ or PIQ > VIQ, this is necessary but not sufficient for individual diagnostic utility, and until such diagnostic utility is demonstrated in differentiating individuals within these groups, clinicians should not assume that a VIQ/VCI and PIQ/POI/PRI difference is an indicator, marker, or sign for that diagnostic group. Study of distinct clinical groups may well reflect the problem of inverse probabilities where members of a distinct group may likely demonstrate VIQ-PIQ "signs" (sensitivity), but those who demonstrate VIQ-PIQ "signs" may not necessarily be members of that distinct clinical group (positive predictive power). Kaufman and Lichtenberger (2006) noted that, "when evaluating V–P differences for individuals instead of groups, extreme caution must be exercised" (p. 316). In the absence of diagnostic utility research affirming the diagnostic utility of scores for individual diagnostic purposes (especially their positive predictive power), interpretation of those scores should probably be curtailed.

FACTOR/BROAD-ABILITY SCORE COMPARISONS

Beginning with the WISC-IV, revised Wechsler scales no longer provide VIQ and PIQ scores and now only report factor index scores as Stratum II abilities, as they are more factorially pure indexes of latent abilities. Like Full Scale scores, interpretation of factor or broad-ability scores typically first involves a presentation and description of the standard score, percentile rank, and confidence interval (obtained score or estimated true score) for each standard score to account for measurement error (AERA, APA, NCME, 1999). Classification of factor or broad-ability scores within a descriptive category (or range of descriptive categories around

the confidence interval) reflecting some deviation from average is also frequently made (i.e., "average," "below average," "significantly above average"). Like Full Scale scores, these normative descriptors are intended to provide an illustration of how the individual performed relative to others their age and are a function of the interval level measurement the standardized scores represent.

Because there are multiple factor or broad-ability scores, test authors and publishers provide in their respective manuals procedures for comparing these scores to each other. Tables of critical values of difference scores as well as base rates for differences are presented in test manuals and provide clinicians a convenient way to determine which factor or broad-ability scores differ and how rare such a difference was in the standardization sample. Like VIQ-PIQ differences, the inference from significant differences between factor or broad-ability scores is that the individual has stronger cognitive skills in one area than the other and there is speculation as to the implications of these strengths and weaknesses. Factor or broad difference scores that are not sufficiently reliable cannot be valid or of value. Like the VIQ and PIQ scores, factor or broad area scores are not independent and inferences from them are also obscured by shared variance.

FACTOR/BROAD-ABILITY PSYCHOMETRIC SUPPORT

Psychometrically, factor scores or broad-ability scores typically have internal consistency estimates, short-term stability estimates, and predictive validity coefficients that are generally lower than the Full Scale score but higher than individual subtest scores (Bracken & McCallum, 1998b; Elliott, 2007b; Glutting, Adams, & Sheslow, 2000b; Kaufman & Kaufman, 1983, 1993, 2004a, 2004b; Naglieri & Das, 1997b; The Psychological Corporation, 1999; Reynolds & Kamphaus, 2003; Roid, 2003b; Wechsler, 1997, 2002a, 2002b, 2003, 2008b; Wechsler & Naglieri, 2006; Woodcock et al., 2001). This is expected, as true score theory predicts that scores with more items and subtests will have less error variance and thus greater reliability and true score variance. This also means that factor scores or broad-ability scores typically include more error variance than the Full Scale score. However, some factor scores or broad-ability scores have better reliability estimates than others, partly related to the number of subtests (and items) that comprise the factor-based score. In long-term stability studies of the WISC-III (Canivez & Watkins, 1998, 1999, 2001) for example, it was found that across the total sample and across age, sex, race or ethnicity, and disability groups that the VCI, POI, FDI, and PSI scores had lower stability coefficients than the FSIQ but more importantly, only the VCI and POI scores showed long-term temporal stability coefficients close to being high enough for individual interpretation or decision-making ($r \geq .85$; Hills, 1981; $r \geq .90$; Aiken, 2000; Guilford & Fruchter, 1978; Nunnally & Bernstein, 1994; Ponterotto & Ruckdeschel, 2007; Salvia & Ysseldyke, 1988, 2001). The FDI and PSI score stability coefficients were too low (too unstable) for individual clinical use. Similar results were also obtained by Krohn and Lamp (1999) for the K-ABC and SB:FE.

FACTORIAL/STRUCTURAL VALIDITY

While some factor-based scores might possess acceptable reliability coefficients (internal consistency, short-term stability, long-term stability) and reliability is a foundation for possible score validity, validity is ultimately more important. Also, "validity is always *specific to some particular use or interpretation*" (Linn & Gronlund, 1995, p. 49). Investigations of the internal or structural validity of intelligence tests is often conducted via factor analyses (exploratory [EFA] *and* confirmatory [CFA]), but recently, some intelligence test authors and publishers (Elliott, 2007b; Roid, 2003b, Wechsler, 2008b; McGrew & Woodcock, 2001) opted to report only results from CFA. This is in contrast to previous practice (and some current practice) wherein both EFA *and* CFA results were both reported (Bracken & McCallum, 1998b; Elliott, 1990; Glutting et al., 2000b; Kaufman & Kaufman, 1993; Naglieri & Das, 1997b; Wechsler, 1991, 2002a, 2002b; Wechsler & Naglieri, 2006). Gorsuch (1983) noted the complimentary nature of EFA and CFA, suggesting that greater confidence in the internal structure of a test is obtained when EFA and CFA are in agreement. As noted by Frazier and Youngstrom (2007), there is good cause for concern regarding the disagreement between the number of latent factors reported in contemporary intelligence tests based only on CFA procedures (or the most liberal EFA factor-extraction criteria) and the number of factors suggested with EFA procedures using the most psychometrically sound methods for determining the correct number of factors to extract and retain. For example, DiStefano and Dombrowski (2006) and Canivez (2008) provided markedly different results for the SB-5 than the CFA results presented in its technical manual (Roid, 2003b).

Another EFA approach to investigate the internal structure of intelligence tests is the Schmid and Leiman (1957) procedure, which was recommended by Carroll (1993, 1995, 1997a, 2003); McClain (1996); Gustafsson and Snow (1997); Carretta and Ree (2001); Ree, Carretta, and Green (2003); and Thompson (2004). Because the narrow abilities (subtests) and broad abilities (factors) are themselves correlated, subtest performance on cognitive abilities tests reflect combinations or mixtures of both first-order *and* second-order factors. Carroll argued that variance from the second-order factor should be extracted first to residualize the first-order factors, leaving them orthogonal to each other and the second-order factor. Thus, variability associated with the higher-order factor is accounted for prior to interpreting variability in the lower-order factors. In this way, it is possible to see how the reliable test variance is partitioned to higher- and lower-order dimensions. However, almost no test manuals provide these analyses for practitioners to review.

When the Schmid and Leiman (1957) procedure has been used with contemporary intelligence tests, the higher-order factor (*g*) accounted for the largest portion of variance, and considerably smaller portions of variance remained at the lower-order (factors) level (Bracken & McCallum, 1998b; Canivez, 2008, 2011; Canivez, Konold, Collins, & Wilson, 2009; Canivez & Watkins, 2010a, 2010b; Domrowski & Watkins, in press; Dombrowski, Watkins, & Brogan, 2009; Nelson & Canivez, 2012; Nelson, Canivez, Lindstrom, & Hatt, 2007; Watkins, 2006; Watkins, Wilson, Kotz, Carbone, & Babula, 2006). This is one reason the primary (if not exclusive) interpretation should be at the Full Scale score level. Clinicians *should be* provided such information about the portions of test variance captured at the different levels of the test in test manuals to facilitate decisions about the importance of the different dimensions and what should be interpreted. Unfortunately, this information is absent from most contemporary intelligence test technical manuals. However, decisions about the validity and interpretation of intelligence tests cannot be sufficiently answered or resolved using only structural validity or internal structure perspective (EFA or CFA) (Canivez et al., 2009; Carroll, 1997b; Kline, 1994; Lubinski & Dawis, 1992).

FACTOR/BROAD-ABILITY INCREMENTAL VALIDITY

When considering intelligence test validity and interpretation across multiple levels and scores from a test, it is critical to consider the *external* validity investigations such as predictive validity and incremental validity of lower-level scores beyond that of higher-level scores (Haynes & Lench, 2003; Hunsley, 2003; Hunsley & Meyer, 2003). In this way the relative importance of factor scores versus the global Full Scale score may be assessed. However, *validity* should not be confused with *diagnostic utility* (Meehl, 1959; Mullins-Sweatt & Widiger, 2009; Wiggins, 1988), as the latter is concerned with the application of test score interpretation to the individual. It follows that construct validity and criterion-related validity, Cronbach and Meehl (1955) preferred construct validity, are a prerequisite without which utility is not possible.

A major aspect of intelligence test use is its utility in assisting in the diagnosis or classification of an individual (e.g., MR, SLD, GT). Examination of the diagnostic or predictive utility is also a prerequisite for the ethical use of test scores (Dawes, 2005). Ultimately, the greatest utility would be the ability of a test or set of variables to accurately determine the likelihood of treatment response under specified conditions (i.e., treatment validity). However, prediction is, in and of itself, important, regardless of treatment utility (Glutting, Watkins, & Youngstrom, 2003) and is frequently investigated.

The importance of incremental validity investigations in general, and in the case of multilevel intelligence test interpretation in particular, is based on an important scientific principle articulated by William of Ockham (alt. "Occam"): the law of parsimony (Occam's razor), which states "what can be explained by fewer principles is needlessly explained by more" Jones, 1952, p. 620). Thus, science favors a less complex explanation over a more complex explanation for phenomena. In the case of intelligence test interpretation, the Full Scale score, an estimate of *g*, is a more parsimonious index than the lower-level factor or broad-ability scores (and subtest scores) and satisfies the law of parsimony. Intelligence test Full Scale scores demonstrate substantial criterion-related validity (Neisser et al., 1996; Carroll, 1993; Gottfredson, 1997, 2008, 2009; Jensen, 1998; Lubinski, 2000; Lubinski & Humphreys, 1997), so in order for the factor scores to be relevant, they must demonstrate *meaningful* predictive validity *beyond* that afforded by the Full Scale score.

Besides describing performance on factor-based scores or broad-ability scores, clinicians are often instructed to consider predictive utility and explanation of performance in academic achievement areas reflecting the higher- and lower-order factor scores.

For example, if significant differences between factor scores exist, or if factor scores deviate from the individual's mean factor performance, that variability among factor scores suggests to some that the FSIQ is not interpretable and that the clinician must examine and interpret the examinee's unique pattern of performance on the factors or broad abilities (Flanagan & Kaufman, 2004; Hale & Fiorello, 2004; Kaufman, 1994; Kaufman & Lichtenberger, 2002, 2006; Lezak, 1995; Sattler, 2008; Sattler & Ryan, 2009; Weiss, Saklofske, & Prifitera, 2003; Wolber & Carne, 2002). Others (Gridley & Roid, 1998; Hale & Fiorello, 2004; Hildebrand & Ledbetter, 2001) have suggested that under these conditions the FSIQ would not be a valid predictor of the individual's academic achievement. Even when there are no differences among factor scores, interpretation of the factor scores may still be done. Those promoting the clinical approach to test interpretation of factor index variability often argue that, while the deviations might not be appropriate for diagnosis, the ability patterns (strengths and weaknesses) could be helpful for instructional strategies, interventions, or treatments or provide hypotheses about the individual (Flanagan & Kaufman, 2004; Kaufman, 1994; Kaufman & Lichtenberger, 2002, 2006; Sattler, 2008; Sattler & Ryan, 2009).

Whether or not such factor score differences provide useful indications for treatment, accommodations, or hypothesis-generation will, in part, be based upon their incremental validity. A primary use of intelligence tests is to predict or account for academic achievement, and if the index scores are to be of practical clinical utility, they must account for meaningful portions of achievement variance *beyond* that provided by the Full Scale score (Haynes & Lench, 2003; Hunsley & Meyer, 2003). This is a necessary, but not sufficient, condition for clinical utility and use with *individuals*. In considering incremental validity, there are two approaches that are often taken and are highly dependent upon the nature of the question being asked and the level of analysis.

In their innovative and highly influential article "Distinctions Without a Difference:…," Glutting, Watkins, Konold, and McDermott (2006) thoroughly examined the validity of observed scores *and* latent factors from the WISC-IV in estimating reading and mathematics performance on the WIAT-II using the WISC-IV—WIAT-II standardization linking sample. Both approaches are important and legitimate methods, but they answer different questions and use different statistical procedures. If one is interested in testing theory and *explaining* latent achievement constructs from latent intelligence constructs, then the use of structural equation modeling (SEM) is an appropriate statistical method (Glutting et al., 2006). However, because the latent constructs are not directly observable, *and* latent construct scores are difficult to calculate and not readily available, there are no direct practical clinical applications (Oh, Glutting, Watkins, Youngstrom, & McDermott, 2004). If one is interested in clinical application of test scores in *predicting* academic achievement from intelligence test scores, hierarchical multiple regression analysis (HMRA) is an appropriate statistical method (Glutting et al. 2006) and may be the most common statistical method in incremental validity (McFall, 2005). HMRA techniques utilize the observed IQ and factor scores that psychologists have available to them. Unlike the perfectly reliable latent constructs in SEM, in clinical assessment and interpretation for individuals, psychologists *must* use observed scores from tests, and those scores contain measurement error.

Glutting et al. (2006) demonstrated that the WISC-IV FSIQ predicted substantial portions of variance in reading and mathematics scores on the WIAT-II, but the four factor index scores (VCI, PRI, WMI, PSI) did not contribute additional meaningful prediction beyond the FSIQ. Other studies of incremental predictive validity before and after Glutting et al. (2006) provided similar results (Canivez, 2011; Freberg, Vandiver, Watkins, & Canivez, 2008; Glutting, Youngstrom, Ward, Ward, & Hale, 1997; Kahana, Youngstrom, & Glutting, 2002; Ryan, Kreiner, & Burton, 2002; Watkins, Glutting, & Lei, 2007; Youngstrom, Kogos, & Glutting, 1999). Glutting et al. concluded that their results were very like "previous epidemiological studies from both the United States and Europe that showed specific cognitive abilities add little or nothing to prediction beyond the contribution made by *g* (Jencks et al., 1979; Ree, Earles, & Treachout, 1994; Salgado, Anderson, Moscoso, Bertua, & de Fruyt, 2003; Schmidt & Hunter, 1998; Thorndike, 1986)" (2006, p. 110).

Furthermore, in their SEM analyses, Glutting et al. (2006) found only the higher-order *g* and the VC latent construct offered significant *explanations* of reading and mathematics constructs (PR, WM, and PS constructs provided no increases in explanation). Similar SEM findings were also reportedly obtained with the Woodcock-Johnson Revised Tests of Achievement (WJ-R; Woodcock & Johnson, 1989) (Keith, 1999; McGrew, Keith, Flanagan,

& Vanderwood, 1997) and with the WISC-III (Wechsler, 1991; Oh et al., 2004). Kuusinen and Leskinen (1988) and Gustafsson and Balke (1993) reportedly reached similar conclusions with other measures of ability and achievement (Glutting et al., 2006).

Why the factor scores failed to add incremental predictive validity over and above the FSIQ may relate to the earlier discussion of hierarchical EFA where the lower-order factors accounted for substantially smaller portions of reliable variance (Canivez, 2008, 2011; Canivez & Watkins, 2010a, 2010b; Nelson & Canivez, 2012; Nelson et al., 2007; Watkins, 2006; Watkins et al., 2006). If test authors and publishers are interested in improving the incremental predictive validity of cognitive tests, it may be necessary to (a) increase the number of subtests estimating the factor scores to capture more variance, and/or (b) construct cognitive subtests that contain less *g* variance (and more Stratum II or broad-ability variance). However, at present, empirical results continue to corroborate the overwhelming majority of the reliable criterion variable variance is predicted by the single Full Scale intelligence test score (Thorndike, 1986).

Multiple regression analysis research with the WJ-III cognitive clusters predicting reading (Evans, Floyd, McGrew, & Leforgee, 2001) and writing (Floyd, McGrew, & Evans, 2008) found some clusters were more important than others. However, these were not *hierarchical* multiple regression analyses first accounting for *g* and then accounting for cluster score improvement in predicting academic achievement. Thus, the incremental validity of clusters beyond *g* was not investigated. Other recent WJ-III research used SEM procedures to examine direct vs. indirect *explanations* of *g* with direct vs. indirect *explanations* of broad-ability dimensions in areas of reading decoding (Floyd, Keith, Taub, & McGrew, 2007) and mathematics achievement (Taub, Floyd, Keith, & McGrew, 2008). Both studies noted the WJ-III influences of *g* were large but indirect through the broad-ability dimensions. However, as Floyd et al. noted, the WJ-III has a problem regarding possible criterion contamination that could inflate the predictive power of some broad-ability dimensions. Determining direct vs. indirect influences of general intelligence is further complicated and unresolved due to issues of singularity, multicollinearity, and reported Heywood cases in SEM of the J-III in the Floyd et al. study (i.e., Gf–g, Glr–g, Gsm–g [three-stratum model] and Gf–g [two-stratum model]; paths at 1.0). Another important issue remains, despite these authors' arguments for practitioner use of SEM results in informing test interpretation. Glutting et al. (2006) pointed out,

> We previously demonstrated the following: (a) The constructs from SEM rank children differently than observed scores, and children's relative position on factor-based constructs (e.g., VC) can be radically different than their standing on corresponding observed factor scores (the VCI); (b) construct scores are not readily available to psychologists; and (c) although it is possible to estimate construct scores, the calculations are difficult and laborious (cf. Oh et al., 2004, for an example). Therefore, one of the most important findings here is that psychologists cannot directly apply results from SEM. (p. 111)

Thus, SEM results provide theoretical *explanations* for relationships between the cognitive and achievement variables, but this does not mean that there is direct application in the prediction of achievement performance from the cognitive test scores. Thus, the incremental predictive validity of factor or broad-ability scores for clinical use is very much in doubt.

Perhaps the most extreme view regarding the clinical value of the FSIQ is that of Hale, Fiorello, and colleagues (Fiorello, Hale, Holdnack, Kavanagh, Terrell, & Long, 2007; Fiorello, Hale, McGrath, Ryan, & Quinn, 2001; Hale & Fiorello, 2004; Hale, Fiorello, Bertin, & Sherman, 2003; Hale, Fiorello, Kavanagh, Holdnack, & Aloe, 2007; Hale, Fiorello, Kavanagh, Hoeppner, & Gaither, 2001), who proclaimed the invalidity of the FSIQ in predicting academic achievement when significant intracognitive variability (factor or subtest scatter or variation) was observed. They argued that practitioners should "never interpret the global IQ score if there is significant scatter or score variability" (Hale & Fiorello, 2001, p. 132).

The approach that Hale, Fiorello, and colleagues used to render such a recommendation is that of regression commonality analysis of global and factor-index scores from the WISC-III (Wechsler, 1991), and achievement scores from the Wechsler Individual Achievement Test (WIAT; Wechsler, 1992). Another method (later deemed inappropriate) was their entering factor or broad-ability scores into the first block of hierarchical multiple regression and entering the Full Scale score in the second block to test how much incremental validity there is in the Full Scale score over and above the lower-order factor or broad-ability scores. This approach was

criticized by Glutting et al. (2006), who wrote that while it had "intuitive appeal," and was "employed on occasion (Hale, Fiorello, Kavanagh, Hoeppner, & Gaither, 2001)" (p. 106), such use violates the law of parsimony such that psychologists would favor a more complex accounting for predictive validity rather than the less complex predictor (*g*) when the many factors at best only account for marginally more achievement variance.

A special issue of the journal *Applied Neuropsychology* further addressed these issues and the merits and conclusions of the approach of Hale, Fiorello, and colleagues (Reynolds, 2007). Fiorello et al. (2007) applied regression commonality analysis to WISC-IV factor index scores obtained from the 228 participants previously diagnosed with learning disability (LD), attention deficit–hyperactivity disorder (ADHD) and traumatic brain injury (TBI) from the special-groups data reported in the WISC-IV Technical Manual (Wechsler, 2003b). However, they only included participants with FSIQ scores between 80 and 120 "to ensure extreme scores did not affect study results" (Fiorello et al., 2007, p. 5). Primary conclusions of their results were that the WISC-IV FSIQ is not appropriate for interpretation for these groups (those with intracognitive variability) due to small, shared variance of the four index scores; and individual idiographic interpretation is appropriate based on sizable unique variance components.

The manuscript of Fiorello et al. (2007) was provided to several statistics and psychological measurement experts for critique and comment who provided very different assessments and conclusions. Dana and Dawes (2007); Faust (2007); and Watkins, Glutting, and Lei (2007) pointed out numerous methodological errors as well as empirical evidence arguing against the Hale, Fiorello, et al. use of regression commonality analysis. Hale et al. (2007) provided a rejoinder to address the critiques but appeared to only restate their original position rather than rebut the critiques and data presented (Dana & Dawes, 2007; Faust, 2007; Watkins et al., 2007). Daniel (2007) also provided a critique of Fiorello et al. and used a simulation study to demonstrate that high levels of index-score scatter *did not* affect the FSIQ predictive validity. Schneider (2008), quite dissatisfied with the Hale et al. (2007) rejoinder indicating they did not recognize the flaws in their analyses, provided yet another critique of the Hale, Fiorello, and colleagues' application of regression commonality analysis. Daniel (2009) also provided evidence that WISC-IV subtest or factor score variability does not invalidate the FSIQ in predicting WIAT-II performance, as he showed in comparisons of high- and low-variability groups. In an investigation of the predictive validity of the DAS general conceptual ability index (GCA; Elliott, 1990) when significant and unusual scatter was observed, Kotz, Watkins, and McDermott (2008) found no significant differences in predicting academic achievement by the GCA across groups showing significant *and* clinically unusual differences between factors.

In summary, while factor or broad-ability scores may possess higher reliability estimates than subtest scores, and some have acceptable reliability estimates to support individual decision-making, the validity research does not provide strong enough support for their interpretations in many instances. Furthermore, this discussion was concerned with the issue of statistical incremental validity, not clinical incremental validity, which Lilienfeld, Wood, and Garb (2006) noted could negatively affect decisions based on the "dilution effect," whereby "presenting participants with accurate but nondiagnostic information . . . often results in less accurate judgments" (p. 11), as reported by Nisbett, Zukier, and Lemley (1981). Unless stronger support is provided for their incremental validity, clinicians should restrain their clinical interpretations to the Full Scale score in most, if not all, instances.

Subtest-Based Score Interpretation/ Stratum I

Interpretation of intelligence test subtest scores is most frequently conducted through examination of subtest deviations from the individual's average subtest performance through ipsative comparisons, an ideographic procedure. As noted by Rapaport et al. (1945–1946), the examination of subtest highs and lows (strengths and weaknesses) for an individual was to provide the clinician with valuable information about the individual that could assist in diagnosis and treatment. Zeidner (2001) recommended the use of cognitive strengths and weaknesses derived from the WISC-III as the basis for psychoeducational recommendations. As with other test scores, investigation of reliability and validity of subtest scores is a requirement for determining their utility and thus interpretability.

SUBTEST PSYCHOMETRIC SUPPORT

While Full Scale scores (and some factor or broad-ability scores) demonstrate uniformly high estimates of reliability and validity, the same cannot be said for subtest scores. Great variability

exists within and between various intelligence tests as to the magnitude of their subtest reliability estimates. Invariably, intelligence test subtests typically have lower internal consistency estimates than composite scores (Bracken & McCallum, 1998b; Elliott, 2007b; Glutting et al., 2000b; Kaufman & Kaufman, 1983, 1993, 2004a, 2004b; Naglieri & Das, 1997b; Psychological Corporation, 1999; Reynolds & Kamphaus, 2003; Roid, 2003b; Wechsler, 2002, 2003, 2008b; Wechsler & Naglieri, 2006; Woodcock et al., 2001). Importantly, internal consistency estimates provide the highest estimates of intelligence subtest reliability because they do not consider important sources of error such as temporal stability, scoring errors, or administration errors (Hanna et al., 1981). In examining the long-term stability of WISC-III scores, Canivez and Watkins (1998) found the stability coefficients for subtests ranged from .55 to .75; thus, none showed acceptable stability for individual clinical decision-making. Considering more stringent criteria for reliability estimates for individual clinical interpretation (Aiken, 2000; Hills, 1981; Guilford & Fruchter, 1978; Nunnally & Bernstein, 1994; Ponterotto & Ruckdeschel, 2007; Salvia & Ysseldyke, 1988, 2001), many (most) intelligence test subtests are inadequate. For the subtests with reliability coefficients (internal consistency, short-term stability, long-term stability) that meet or exceed minimum standards, it is also necessary to know how much subtest specificity exists (reliable subtest variance *unique* to that subtest). More importantly, subtest score *validity*, particularly incremental validity, must be empirically supported, or their measurement may simply be redundant.

IPSATIVE SUBTEST COMPARISONS

While interpretation of individual subtest scores in isolation is not very common, the use of intricate subtest interpretation systems (Kaufman, 1994; Kaufman & Lichtenberger, 2000, 2006; Sattler, 2001, 2008; Sattler & Ryan, 2009) is very popular, both in psychology graduate training and in clinical practice (Alfonso et al., 2000; Groth-Marnat, 1997; Kaufman, 1994; Pfeiffer et al., 2000). The argument is that if there is substantial scatter or variability among the subtests, then an IQ score (or factor score) "represents a summary of diverse abilities and does not represent a unitary entity" (Kaufman & Lichtenberger, 2000, p. 424). The specific patterns of subtest scores presumably invalidate global intelligence indices (Groth-Marnat, 1997), and subtest scores and subtest composites become the principal focus of test interpretation. Subtests that are significantly higher or lower than the child's own average (i.e., ipsative comparisons) are deemed strengths or weaknesses; and while some authors (Kaufman & Lichtenberger, 2000; Sattler, 2008; Sattler & Ryan, 2009) point out that such ipsative or subtest comparisons are not diagnostic, they simultaneously claim that such strengths and weaknesses allow the psychologist to formulate hypotheses concerning the underlying problems and implications for the individual. Such hypotheses are then to be examined with other data sources and used for recommending educational or psychological treatment.

If these hypotheses are to be of use, they must be based on scores or results that have acceptable reliability, otherwise one may be formulating hypotheses about characteristics or possible interventions with essentially random indicators. Furthermore, "Any long-term recommendations as to a strategy for teaching a student would need to be based on aptitudes that are likely to remain stable for months, if not years" (Cronbach & Snow, 1977, p. 161). If suggestions regarding differential teaching styles, curricular materials, interventions, and learning environments (Kaufman, 1994; Kaufman & Lichtenberger, 2000; Sattler, 2008; Sattler & Ryan, 2009) are made based on subtest interpretive methods, then investigation of the reliability and validity of such subtest interpretive methods is imperative.

IPSATIVE SUBTEST COMPARISON PSYCHOMETRIC SUPPORT

Watkins (2003) provided a comprehensive and thorough review of the literature regarding intelligence test subtest analyses and noted the overwhelming shortcomings and failures of subtest analyses to reliably and validly inform psychological practice. The temporal stability of WISC-R's (Wechsler, 1974) cognitive strengths and weaknesses was examined by McDermott, Fantuzzo, Glutting, Watkins, and Baggaley (1992), who found that classification stability of the relative cognitive strengths and weaknesses identified by subtest elevations and depressions was near chance levels. Livingston, Jennings, Reynolds, and Gray (2003) also found the multivariate stability of WISC-R subtest profiles across a three-year retest interval too low for clinical use. Watkins and Canivez (2004), in examining WISC-III subtest ipsative strengths and weaknesses and numerous subtest composites across a three-year retest interval, found agreement, on average, at chance levels. Furthermore, *none* of the 66 subtest composites reached the minimum level

of agreement necessary for clinical use (Cicchetti, 1994). Given the poor reliability of ipsative and subtest composite scores, that such scores or profiles would be valid and diagnostically useful is highly unlikely.

Review of the literature on subtest analysis validity and utility (Watkins, 2003; Watkins, Glutting, & Youngstrom, 2005) showed that subtest scores, patterns, and analyses were unable to adequately identify global neurocognitive or neuropsychological deficits presumably related to learning disability (Watkins, 1996), nor were they related to or valid for diagnosis of learning disabilities (Daley & Nagle, 1996; Glutting, McGrath, Kamphaus, & McDermott, 1992; Hale & Raymond, 1981; Hale & Saxe, 1983; Kahana et al., 2002; Kavale & Forness, 1984; Kline, Snyder, Guilmette, & Castellanos, 1992; Livingston et al., 2003; Maller & McDermott, 1997; Mayes, Calhoun, & Crowell, 1998; McDermott & Glutting, 1997; McDermott, Goldberg, Watkins, Stanley, & Glutting, 2006; McGrew & Knopik, 1996; Mueller, Dennis, & Short, 1986; Ree & Carretta, 1997; Smith & Watkins, 2004; Thorndike, 1986; Ward, Ward, Hatt, Young, & Mollner, 1995; Watkins, 1999, 2000, 2003, 2005; Watkins & Glutting, 2000; Watkins & Kush, 1994; Watkins, Kush, & Glutting, 1997a, 1997b; Watkins, Kush, & Schaefer, 2002; Watkins & Worrell, 2000). Furthermore, subtest analyses were not valid in the classification of behavioral, social, or emotional problems (Beebe, Pfiffner, & McBurnett, 2000; Campbell & McCord, 1996, 1999; Dumont, Farr, Willis, & Whelley, 1998; Glutting et al., 1992; Glutting et al., 1998; Lipsitz, Dworkin, & Erlenmeyer-Kimling, 1993; McDermott & Glutting, 1997; Piedmont, Sokolove, & Fleming, 1989; Reinecke, Beebe, & Stein, 1999; Riccio, Cohen, Hall, & Ross, 1997; Rispens et al., 1997; Teeter & Korducki, 1998).

Kaufman (1994) argued that an individual's cognitive pattern "becomes reliable by virtue of its cross-validation" (p. 31) if it is supported by other clinical information and observations. In Kaufman's system, clinicians are thought of as detectives attempting to make sense out of profiles and searching for clues to the individual's strengths and weaknesses within the test and also by supplemental test information (Kaufman & Lichtenberger, 2006). Dawes (1994), however, noted the difficulty (impossibility) of combining different types (and amounts) of information in clinical decision-making, but asserted that the suggestion that unreliable cognitive subtest scores or patterns become valid for the individual when informally and subjectively integrated with a complex mixture of other assessment data simply is not consistent with the empirical literature (Dawes, Faust, & Meehl, 1989). Psychologists are particularly vulnerable to errors in clinical decision-making precisely in situations such as this (Davidow & Levinson, 1993; Faust, 1986, 1990; Watkins, 2003, 2009). Thus, as Faust (1990) noted, the "common belief in the capacity to perform complex configural analysis and data integration might thus be appropriately described as a shared professional myth" (p. 478). Kaufman and Lichtenberger (2006) noted, "The validity that comes from group data may never be available for the individual profile approach that we advocate" (p. 413).

Watkins and Canivez (2004) concluded as follows:

> (a) Recommendations based on unreliable ipsative subtest comparisons or subtest composites must also be unreliable;
>
> (b) Intelligence subtest analysis procedures that lack reliability or agreement across time cannot be valid;
>
> (c) Most students will exhibit several relative cognitive strengths and weaknesses, so their presence should not be interpreted as unusual or pathognomonic;
>
> (d) The fact that several strengths and weaknesses will typically be observed makes it more likely that errors will result from inferring pathology from them; and
>
> (e) Using an essentially random component (i.e., the subtest profile or subtest composite) and then searching for corroborating information, is likely to decrease the accuracy of clinical decision-making.

Meehl and Rosen (1955) noted such impacts in judgement accuracy when attempting to detect low-prevalence strengths or weaknesses. For an elaborative description of the many types of diagnostic decision-making and clinical judgment errors and how clinicians can avoid them, the reader is directed to Watkins (2009), Garb (2005), and Garb and Boyle (2003).

Despite all this negative empirical research, test authors and publishers continue to describe ipsative subtest analysis procedures in test manuals (Bracken & McCallum, 1998b; Elliott, 2007b; Glutting et al., 2000b; Kaufman & Kaufman, 1983, 1993, 2004a, 2004b; Naglieri & Das, 1997b; Reynolds & Kamphaus, 2003; Roid, 2003b; Wechsler, 2002, 2003, 2008b; Wechsler & Naglieri, 2006). Some test authors, however, have

attempted to minimize their use of ipsative subtest comparisons because of their awareness of the lack of empirical support (Glutting et al., 2000b; Reynolds & Kamphaus, 2003). Textbook authors also continue to describe and promote ipsative and subtest composite interpretations (Flanagan & Kaufman, 2004; Kaufman & Lichtenberger, 2000, 2006; Sattler, 2008; Sattler & Ryan, 2009). Continued presentation of such procedures perpetuates the decades-long shared professional myth that such analyses, in the hands of the trained and skilled clinician, provide important clues in understanding the individual examinee. Lilienfeld et al. (2006) presented several reasons why questionable psychological tests remain popular, and two in particular appear to be operating in the domain of ipsative comparisons and profile analyses in intelligence tests. They referred to the belief in special expertise and intuition in combining test scores and other information to render valid interpretations from invalid scores as "the Alchemist's Fantasy," and the influence of "Clinical Tradition and Educational Inertia" also seems to perpetuate these practices. Macmann and Barnett (1997) may well be correct in their characterization of these ipsative subtest interpretations as the "myth of the master detective" (p. 197).

Psychometric Interpretation Conclusion

Each of the psychometric interpretation methods discussed above requires the psychologist to consider the scores and render an inference or decision about the individual based on their judgment. Elliott (2007b) wrote, "Profile interpretation is clinical rather than statistical; suggestive rather than definitive; and concerned with hypothesis generation" (p. 93). However, as Dawes (1994) pointed out, "*The accuracy of the judgment of professional psychologists and other mental health workers is limited, however, by the accuracy of the techniques they employ*" (p. 107). While there is abundant research support for the clinical interpretation of omnibus, Full Scale intelligence test scores, such is not the case for clinical interpretation of factor scores; and especially subtest scores, profiles, or patterns. Clinical interpretation of intelligence test subtests is essentially the interpretation of scores that have too much error for individual use and will lead to significant errors in formulating hypotheses as well as in diagnosis and treatment recommendations. Even factor-based or broad-ability scores are questionable when their incremental predictive validity estimates are unremarkable, as previously illustrated. At present, ample evidence for clinical interpretation of Full Scale scores from intelligence tests exists and should be the primary, if not exclusive, interpretation focus. For those promoting subtest and factor score or broad-ability score interpretations, it is incumbent on them to provide strong empirical evidence for their interpretation procedures, particularly their utility in the correct prediction of diagnostic groups or disorders, and more importantly, differential treatment (McFall, 1991, 2000). At present, such evidence does not exist.

Actuarial Interpretation Methods

Actuarial test interpretation involves a statistically based decision regarding an individual based on scores from one or more measures (one or more variables). Data could include test scores from standardized tests, but also could include ratings, interview information, and historical information. The statistical combination of available data (i.e., logistic regression, discriminant function analysis, multiple regression, etc.) optimizes the prediction. These statistical procedures are able to differentially weight variables in predictions, and only the variables that have significant contribution to prediction are retained and used. Such complex combinations of variables are something clinicians simply are unable to do (Dawes et al., 1989; Faust, 1990). Decisions one might be interested in making about an individual include classification of the individual's profile (i.e., "Which empirically based profile does the individual's profile most resemble; or is it unique?"), diagnostic or classification decisions (i.e., differential diagnosis), or determining the probability of success for a given treatment (i.e., given this individual's characteristics, treatment *x* is expected to produce some likely response). It is sometimes argued that, in order to make an actuarial interpretation of an intelligence test, one must have access to formulae or data that have been developed and (hopefully) cross-validated on a new sample to provide a comparison of an individual's test score(s). Such methods require available outcome data by which one may derive algorithms for comparison.

Over 50 years ago, Paul Meehl set in motion a debate on actuarial prediction (decision making) by seeking answers to questions about the relationship between clinical and actuarial (statistical) prediction in his seminal book, *Clinical versus statistical prediction: A theoretical analysis and review of the evidence* (Meehl, 1954). His self-proclaimed "wicked book" (Meehl, 1979, p. 564) or "disturbing little book" (Meehl, 1986, p. 370) reviewed

and examined the clinical decision-making (prediction) abilities of clinicians versus actuarial/statistical formula–based predictions. Meehl's conclusion was that the actuarial approach was superior and should be used more frequently. Since that time, there have been numerous studies comparing clinical (informal or impressionistic) and actuarial (formal, mechanical, algorithmic) predictive methods, and it has been fairly consistently shown that the actuarial method is as accurate or more accurate than clinical methods (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000). While 8 of the 136 studies in the Grove et al. (2000) meta-analysis showed superiority of the clinical method, 7 of the 8 benefitted from *more* information via clinical interview not made available to the actuarial method. While most studies in the Grove et al. meta-analysis found a statistical equivalence between the clinical and actuarial methods, it has been argued that in the event of a tie, there should be preference for the actuarial method, because once developed it is less expensive in time and money, less laborious, and allows for consistent application in a dispassionate manner (Dawes et al., 1989; Meehl, 1954).

Why might an actuarial/statistical/mechanical method of prediction be superior? The answer appears to be, in part, its consistent application. All one need do is correctly enter the appropriate scores or data into the formula, and the formula calculates the prediction consistently. It has been reported numerous times that humans (expert clinicians included) are susceptible to numerous errors in judgement, including confirmation bias, overconfidence, fundamental attribution error, misperception of regression, representativeness, insensitivity to prior probabilities or base rates, misperception about chance (i.e., illusory correlations, conjunction fallacy, inverse probabilities, insensitivity to sample size [law of small numbers], pseudodiagnosticity), and hindsight bias (Garb, 1997, 1998; Kahneman, Slovic, & Tversky, 1982; Meehl & Rosen, 1955; Tversky & Kahneman, 1974; Watkins, 2009). McDermott (1981) also noted problems such as the inconsistent application of diagnostic criteria (decision rules), inconsistent weighting of diagnostic cues, and inconsistent decision-making processes (strategies or sequences) among school psychologists. However, entering data into formulae in a consistent manner allows the algorithm or calculations to be applied consistently and resulting decisions from them to be applied consistently as well. Another important aspect of actuarial or statistical superiority rests in the variables included in the formula. Statistical methods of multiple regression, logistic regression, and discriminant function analysis are able to differentially and optimally weight variables to provide the most accurate predictions of the criterion variable, and this provides another advantage over that of a clinician (Grove & Meehl, 1996). It is for these and other reasons that Grove and Meehl argued that actuarial methods should be widely applied and false arguments against it should be rejected.

Research on actuarial interpretations of intelligence tests is quite sparse. Literature searches crossing key terms such as *intelligence test, psychometric intelligence, interpretation, actuarial, statistical, classification, diagnosis,* or *prediction* produced no empirical research applied to actuarial intelligence test interpretation. There are, however, some applications and approximations worth examining.

Statistical/Actuarial Approaches: Classification of Intelligence Test Profiles

Intelligence test subtest (or factor score) profile analysis as systematized by Kaufman (Kaufman, 1979, 1994a; Kaufman & Lichtenberger, 2000) and Sattler (1982, 1988, 1992, 2001, 2009; Sattler & Ryan, 2009) is an ideographic method that uses the individual's mean performance as the basis for comparing subtest (or factor) scores, and determination of strengths or weaknesses is based on significant deviation from that mean. As previously reviewed, these ipsative approaches are neither reliable nor valid in distinguishing clinical group memberships. However, another approach to examining subtest profiles in tests is a *normative* method whereby characteristic profiles are identified through procedures such as cluster analysis (Hale, 1981; McDermott, 1998; Ward, 1963).

Several methods of cluster analysis are available and involve examining individuals' scores on a test and grouping similarly scoring individuals into mutually exclusive groups or clusters with a minimal loss of information. McDermott (1998) developed a three-stage hierarchical clustering method, *Multistage Euclidean Grouping* (MEG), which incorporated recommended cluster analysis techniques such as application of Ward's (1963) method (e.g., Konold, Glutting, McDermott, Kush, & Watkins, 1999), combining hierarchical and nonhierarchical clustering algorithms, and included built-in replications (Milligan & Hirtle, 2003). Once clusters are identified, they are then examined for characteristics (internal and external) that deviate from

other clusters' in order to describe distinguishing characteristics. Clusters may differ in proportions of demographic characteristics such as sex, race or ethnicity, and socioeconomic status (SES), as well as performance or scores on other measures (achievement, learning behaviors, personality, psychopathology). When an individual's test scores are compared to the various profiles defined by the clusters, their profile might be assigned to a particular cluster based on similarity, or perhaps the individual has scores that reflect similarity to no other profile, in which case the profile is deemed unique.

A number of intelligence tests have been examined through cluster analysis in order to determine what profiles exist from a normative perspective. Cronbach and Gleser (1953) noted that profiles are defined by three characteristics (a) level/elevation (i.e., average performance), (b) shape/pattern (i.e., highs and lows or peaks and valleys), and (c) scatter/variability (i.e., range of scores); and profile shape/pattern is determined after removing the level and scatter information. Tests such as the WPPSI, WISC-R, WISC-III, WAIS-R, DAS, UNIT, KABC, and McCarthy Scales of Children's Abilities (MSCA; McCarthy, 1972) have had their standardization samples subjected to cluster analysis and resulting normative profiles described (Donders, 1996; Glutting, & McDermott, 1990a, 1990b; Glutting, McDermott, & Konold, 1997; Glutting, McGrath, Kamphaus, & McDermott, 1992; Holland & McDermott, 1996; Konold et al., 1999; McDermott, Glutting, Jones, & Noonan, 1989; McDermott, Glutting, Jones, Watkins, & Kush, 1989; Schinka & Vanderploeg, 1997; Wilhoit & McCallum, 2002). In all of these examples, the primary distinguishing feature appears to be that of profile level/elevation, which is a reflection of overall ability (*g*). The next distinguishing characteristic of normative profiles appears to be shape/pattern, which often is reflected by broad differences between the test's verbal/crystallized and nonverbal/fluid/visual tasks.

What a normative typology based on cluster analysis affords is a means by which an individual's profile may be *empirically* compared and classified and in a manner that does not discard reliable test variance like the ipsative subtest profile method does (Jensen, 1992; McDermott et al., 1992). Also, group similarity coefficient statistics, such as $r_{p(k)}$ (Tatsuoka, 1974, p. 31; Tatsuoka & Lohnes, 1988, pp. 377–378) or D^2 (Cronbach & Gleser, 1953; Osgood & Suci, 1952), provided an index of similarity to the normative profile types that account for all three profile characteristics. If, for example, none of the normative core profile type comparisons produces an $r_{p(k)}$ value > .40 (Konold et al., 1999; McDermott, Glutting, Jones, Watkins, et al., 1989; McDermott, Glutting, Jones, & Noonan, 1989), then the individual's profile was classified as unique or atypical. Another method of profile comparison is based on Euclidian distance or generalized distance theory (D^2) (Osgood & Suci, 1952), and although somewhat less precise, it is easier to calculate and apply and thus more convenient.

These nonlinear multivariate profile analysis methods are better than clinically based ipsative methods in that they simultaneously consider both linear and nonlinear characteristics of the profile, simultaneously examine multiple subtest scores, and empirically determine similarity (or uniqueness) to the normative core profiles from a nationally representative sample. However, like other test scores, profile similarity or classification must also demonstrate acceptable reliability, validity, and utility.

CLUSTER COMPARISON PSYCHOMETRIC SUPPORT

While normative core profiles have been (or can be) developed for intelligence tests, the measurement properties of the profiles need to be investigated, as well as the measurement properties of individuals' profiles. It was earlier shown that ipsative subtest profiles (strengths and weaknesses) and subtest composite scores were not stable across time and therefore could not be (and were not) valid.

Short-term stability of profile classifications has yielded fairly consistent results for the MSCA (general κ_m =.728; Glutting & McDermott, 1990a), K-ABC (general κ_m =.497; Glutting et al., 1992), and DAS (general κ_m =.541; Holland & McDermott, 1996). Partial κ_m coefficients were also found to be statistically significant for MSCA core profiles (Glutting & McDermott, 1990a) as well as for K-ABC core profiles and a group of unusual K-ABC profiles (Glutting et al., 1992). WPPSI profile short-term stability was lower (general κ_m =.216; Glutting & McDermott, 1990b).

While short-term stability for empirically based profiles was moderate, Livingston et al. (2003) found that empirically derived subtest profiles did not possess acceptable long-term stability; however, they did not evaluate profile stability by comparison to the core taxonomy. Borsuk, Watkins, and Canivez (2006) explored the long-term stability of WISC-III cluster membership based on nonlinear multivariate profile analysis for 585 students across

a mean retest interval of 2.82 years. Individual profiles at Time 1 and Time 2 were classified according to the normative core WISC-III profiles (Konold et al., 1999) using D^2 (Cronbach & Gleser, 1953; Osgood & Suci, 1952) and the critical D^2 value of 98 established by Konold et al. Agreement for all profile types across time (κ_m =.39, p <.0029; Fleiss, 1971) and partial κ_m coefficients for each individual profile (.26 to .51) indicated that cluster membership based on nonlinear multivariate profile analysis was generally not sufficiently stable over a three-year period showing generally poor agreement (Cicchetti, 1994). Profiles 6 and 8 showed fair and statistically significant stability necessary to justify future validity research (Cicchetti, 1994). Although it appears that several intelligence test profile–type memberships possess some degree of short-term stability, long-term stability results for the WISC-III (Borsuk et al., 2006) were generally poor. As such, even the empirically based WISC-III subtest profile-type memberships were not suitable for making educational decisions about students. Thus, at this point, both nonlinear multivariate *and* clinical (ipsative) approaches to profile analysis lack empirical support for contribution to individual diagnosis or educational decision-making.

If empirically derived profiles were at some point found to be reliable, they then must also provide incremental validity over and above general intelligence scores *and* must assist in diagnostic utility for clinical use. However, like ipsative subtest interpretive methods, normative approaches to subtest interpretation have inadequate empirical support in their diagnostic utility (Glutting et al., 1992; Glutting, McDermott, Konold, Snelbaker, & Watkins, 1998; McDermott et al., 1992), and it is appropriate to heed the recommendation, even for normatively based profiles, to "just say no" (McDermott, Fantuzzo, & Glutting, 1990) to all subtest analyses and interpretations in clinical practice.

Statistical/Actuarial Approaches to Classification and Diagnosis

Ultimately, actuarial (statistical) classification and diagnosis should include co-normed measures assessing relevant domains (intelligence, academic achievement, adaptive behavior, personality, learning behaviors, psychopathology) and include large, demographically representative standardization samples. This would allow the generation of multivariate statistical comparisons and enable empirical classification and differential diagnosis. With a demographically representative sample, base-rate estimates for the population would be available for empirically delineated and defined pathologies. This, however, is not yet available.

One group of instruments that is an approximation is the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993), the Learning Behaviors Scale (LBS; McDermott, Green, Francis, & Stott, 1999), and the DAS (Elliott, 1990). In the nationally representative standardization of ASCA by The Psychological Corporation, 1,260 of the 1,400 youths in the ASCA standardization sample were also administered the DAS, and 1,252 had teacher ratings on the LBS. In the cluster analysis of the ASCA (McDermott, 1993, 1994; McDermott & Weiss, 1995) 22 distinct profiles (14 major types, 8 clinical subtypes) were identified based on the six ASCA core syndromes. Following identification of distinct profiles, McDermott and Weiss were able to describe the characteristics of cluster profile types according to features that differed significantly from other profile types on demographic variables (age, sex, SES, race/ethnicity), cognitive abilities (general conceptual, verbal, nonverbal), academic achievement (word reading, numerical skills, spelling), and ability–achievement discrepancies. Using generalized distance scores (D^2), an individual's core syndrome profile is classified as most similar to one of the 22 normative profiles producing the lowest D^2 value and characteristics of the profile likely related to the youth in question. While this application is not directed at intelligence per se, a similar procedure could provide greater understanding and empirically based classification. This co-normed set of tests also allowed McDermott et al. (2006) to examine aspects of intelligence, processing speed, classroom learning behaviors, problem behaviors, and demographic variables in identifying differential risk of learning disabilities from an epidemiological perspective. Important differences were identified in differential risk and classification depending on some of these variables as well as the method of determining learning disability (low achievement vs. ability–achievement discrepancy).

SYSTEMS ACTUARIAL CLASSIFICATION

Recognizing the problem of the lack of consistency or agreement among (and within) diagnosticians in child clinical psychology and school psychology diagnostic decision-making, McDermott (1980) developed a multidimensional system for the actuarial differential diagnosis of children with disabilities. This multidimensional actuarial classification

(MAC) was the forerunner of the McDermott Multidimensional Assessment of Children program (M-MAC; McDermott & Watkins, 1985). Implicit in this process is the notion that there must be reliable application of diagnostic criteria and consideration of multivariate analyses. Without reliability in clinical decision-making, there can be no validity. The M-MAC was generations ahead of its time in terms of both technology and comprehensive actuarial classification. Sadly, nothing like it even exists today!

M-MAC (and its predecessor MAC) applied a classification system that considered both abnormal *and* normal development and provided classifications based, in part, on objective measures of intelligence, academic achievement, adaptive behavior, and psychopathology; recognizing that variations within and between these would provide for differential classification or diagnosis. Because the diagnostic decision rules and mathematical comparisons are applied consistently, the classifications across similar or identical cases are reliable. This is a necessary first step for any method of diagnosis. With respect to intelligence test interpretation, M-MAC provided differential diagnosis for mental retardation (e.g., both intelligence *and* adaptive behavior measures were at least two standard deviations below the mean) and learning disabilities (e.g., IQ–achievement discrepancy, consideration of significant and rare achievement problems, and absence of mental retardation, sensory impairment, etc.). Prevalence rates were also applied and increased the validity of classifications (Glutting, 1986a). M-MAC also provided for the development of recommended intervention programs (1,111 specific behavioral objectives in reading, math, learning, and/or adaptive skills) to address the previously identified diagnostic needs of the child (Glutting, 1986b).

Evaluation of MAC (McDermott, 1980; McDermott & Hale, 1982) with a sample of 73 youths referred to an outpatient clinic resulted in agreement across areas (mental retardation, specific learning disability, behavioral/emotional disorder, communication/perceptual, reading problem, mathematics problem) 86% beyond chance when MAC results were compared to experts'. Agreement between two experts for cases averaged 76.5% beyond chance, but experts were not significantly in agreement for classifications of learning disability or mathematics problems. Agreement for MAC (expert applied vs. novice applied) across disorders was perfect! Thus, as observed elsewhere (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Grove et al., 2000), actuarial interpretations by MAC were generally better than those of experts. Actuarial classification or diagnosis by programs like MAC or M-MAC provide advantages of basing diagnostic decisions on data (i.e., test scores) that are standardized and normed on representative samples, mathematical comparisons of scores from different tests, and statistical decision rules and application of diagnostic criteria are applied consistently and in accordance with diagnostic standards established by governmental agencies or professional organizations.

Actuarial Interpretation Conclusion

Research regarding the benefits of actuarial methods for classification and diagnosis has been available for some time, yet clinical application in interpretation of intelligence and other tests has generally failed to capitalize on this. Actuarial interpretation affords systematic and reliable application and ability for multivariate consideration of variables that affect reliable and valid differential diagnosis. Where the M-MAC program from the 1980s required numerous 5.25" floppy disks to be swapped during computations, the technological advances in computer processing power and storage capacity would allow the complex algorithms to be run on today's handheld computers and smart phones. Better clinical practice and more ethical practice would be afforded by actuarial and empirical applications in intelligence test interpretation.

A final note regarding actuarial methods is necessary concerning their limits (Dawes et al., 1989). An actuarial method is only as good as the measures included for predictions and the available outcomes. Research and evaluation of the actuarial algorithms and accuracy of decisions is necessary in order to continually improve. Quality control and revision based on theory development and research are most certainly necessary as the field advances. Also, actuarial methods should not be considered infallible, as there will always be errors in diagnostic decision-making and within psychology, and because there appear to be no biological or positive markers for disorders, we never really know with certainty whether or not individuals have a particular disorder. As such, clinicians must cope with the reality of practicing with uncertainty.

General Conclusion

Interpretation of intelligence tests requires careful consideration of the empirical support for their reliability, validity, and utility. At present there appear

to be no specific actuarial systems of intelligence test interpretation leading to prediction or differential diagnosis. Available research, considered in its entirety, suggests that most, if not all, interpretation should be based on the overall, or omnibus score (FSIQ). This is not to say that cognitive or intellectual abilities are only one thing (*g*), but at present our ability to measure more than general intellectual abilities is less than adequate when considering important uses such as in prediction of academic achievement and diagnostic decision-making. The inadequacies of lower-level scores beyond the Full Scale score will lead to greater errors in diagnostic decision-making and treatment recommendations. This research has been available for decades, yet numerous test authors and publishers, textbook authors, and university trainers of clinicians continue to perpetuate the clinical interpretation method and the shared professional myth of the utility of subtest and other interpretations and often ignore this research altogether. It is hoped that a new generation of psychologists will heed the empirical evidence and advice of those who have repeatedly called for abandonment of subtest interpretations. It is time to follow Weiner's (1989) sage advice that effective psychodiagnosticians:

> (a) know what their tests can do and (b) act accordingly. Knowing what one's test can do—that is, what psychological functions they describe accurately, what diagnostic conclusions can be inferred from them with what degree of certainty, and what kinds of behavior they can be expected to predict—is the measure of a psychodiagnostician's competence. Acting accordingly—that is, expressing only opinions that are consonant with the current status of validity data—is the measure of his or her ethicality. (p. 829)

However, in clinical assessment, intelligence is but one domain to be considered, and any consideration of multiple domains simultaneously requires multivariate analyses that tax human information processing and clinical judgement. Tests covering many important domains (intelligence, achievement, personality, psychopathology, adaptive behavior, learning behaviors) simultaneously normed on representative population samples would help us improve differential diagnosis and better understand psychopathology base rates, and allow for actuarial interpretation for individual examinees. In the absence of such an ambitious venture, perhaps one day soon a systems-actuarial interpretation program like M-MAC will be created to account for the multivariate measurement of psychopathologies, of which intelligence is one part, to improve the clinical decision-making process when questions of intelligence confront the clinician. Such a method would at least assure that diagnostic criteria would be consistently applied. Only then will there be the possibility of valid classification and diagnosis.

Author Note

Gary L. Canivez is Professor of Psychology at Eastern Illinois University, principally involved in training school psychologists. His research interests include applied psychometric investigations of the reliability, validity, and utility of intelligence, achievement, and psychopathology measures; and investigations of test bias.

The author would like to thank Drs. Marley W. Watkins, W. Joel Schneider, Scott O. Lilienfeld, and Peter V. W. Hartmann for extremely helpful critiques, comments, and suggestions concerning earlier versions of this chapter.

Correspondence regarding this manuscript should be addressed to Gary L. Canivez, Ph.D., Department of Psychology, 600 Lincoln Avenue, Charleston, Illinois 61920-3099. Dr. Canivez may also be contacted via email at glcanivez@eiu.edu, glcanivez@gmail.com, or the World Wide Web at http://www.ux1.eiu.edu/~glcanivez.

References

Aiken, L. R. (2000). *Psychological testing and assessment* (10th ed.). Needham Heights, MA: Allyn & Bacon.

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review*, *29*, 52–64.

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.

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

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

American Psychological Association. (2002, 2010 Amendments). *Ethical principles of psychologists and code of conduct.* Washington, DC: APA.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, *53*, 370–18. (Reprinted in G. A. Barnard [1958], *Studies in the history of probability and statistics*.) *Biometrika*, *45*, 293–315.

Beebe, D. W., Pfiffner, L. J., & McBurnett, K. (2000). Evaluation of the validity of the Wechsler Intelligence Scale for Children–Third Edition comprehension and picture arrangement subtests as measures of social intelligence. *Psychological Assessment*, *12*, 97–101.

Binet, A., & Simon, T. (1905). Methodes nouvelle pour le diagnostic du niveau intellectual des anormaux [New methods for the diangnosis of the intellectual level of subnormals]. *L'Annee Psychologique*, *11*, 1991–244.

Bornstein, R. A., & Matarazzo, J. D. (1982). Wechsler VIQ versus PIQ differences in cerebral dysfunction: A literature review with emphasis on sex differences. *Journal of Clinical Neuropsychology*, *4*, 319–334.

Bornstein, R. A., & Matarazzo, J. D. (1984). Relationship of sex and the effects of unilateral lesions on the Wechsler intelligence scales: Further considerations. *Journal of Nervous and Mental Disease*, *172*, 707–710.

Borsuk, E. R., Watkins, M. W., & Canivez, G. L. (2006). Long-term stability of membership in a WISC-III subtest core profile taxonomy. *Journal of Psychoeducational Assessment*, *24*, 52–68.

Bracken, B. A., & McCallum, R. S. (1998a). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside Publishing.

Bracken, B. A., & McCallum, R. S. (1998b). *Universal Nonverbal Intelligence Test: Examiners manual*. Itasca, IL: Riverside Publishing.

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests for preschool children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 484–502). New York: The Guilford Press.

Brody, N. (1985). The validity of tests of intelligence. In B. Wolman (Ed.), *Handbook of intelligence* (pp. 353–389). New York: Wiley.

Brody, N. (2002). *g* and the one–many problem: Is one enough? In *The nature of intelligence* (Novartis Foundation Symposium 233) (pp. 122–135). New York: Wiley.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since bias in mental testing. *School Psychology Quarterly*, *14*, 208–238.

Campbell, J. M., & McCord, D. M. (1996). The WAIS-R comprehension and picture arrangement subtests as measures of social intelligence: Testing traditional interpretations. *Journal of Psychoeducational Assessment*, *14*, 240–249.

Campbell, J. M., & McCord, D. M. (1999). Measuring social competence with the Wechsler picture arrangement and comprehension subtests. *Assessment*, *6*, 215–223.

Canivez, G. L. (2008). Orthogonal higher-order factor structure of the Stanford-Binet Intelligence Scales for children and adolescents. *School Psychology Quarterly*, *23*, 533–541.

Canivez, G. L. (2011). Hierarchical factor structure of the Cognitive Assessment System: Variance partitions from the Schmid-Leiman (1957) procedure. *School Psychology Quarterly*, *26*, 305–317.

Canivez, G. L. (2011, August). *Interpretation of cognitive assessment system scores: Considering incremental validity of PASS scores in predicting achievement.* Paper presented at the 2011 Annual Convention of the American Psychological Association, Washington, DC.

Canivez, G. L., Konold, T. R., Collins, J. M., & Wilson, G. (2009). Construct validity of the Wechsler Abbreviated Scale of Intelligence and Wide Range Intelligence Test: Convergent and structural validity. *School Psychology Quarterly*, *24*, 252–265.

Canivez, G. L., Neitzel, R., & Martin, B. E. (2005). Construct validity of the Kaufman Brief Intelligence Test, Wechsler Intelligence Scale for Children–Third Edition, and Adjustment Scales for Children and Adolescents. *Journal of Psychoeducational Assessment*, *23*, 15–34.

Canivez, G. L., & Watkins, M. W. (1998). Long term stability of the Wechsler Intelligence Scale for Children–Third Edition. *Psychological Assessment*, *10*, 285–291.

Canivez, G. L. & Watkins, M. W. (1999). Long term stability of the Wechsler Intelligence Scale for Children–Third Edition among demographic subgroups: Gender, race, and age. *Journal of Psychoeducational Assessment*, *17*, 300–313.

Canivez, G. L., & Watkins, M. W. (2001). Long term stability of the Wechsler Intelligence Scale for Children–Third Edition among students with disabilities. *School Psychology Review*, *30*, 438–453.

Canivez, G. L., & Watkins, M. W. (2010a). Investigation of the factor structure of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV): Exploratory and higher-order factor analyses. *Psychological Assessment*, *22*, 827–836.

Canivez, G. L., & Watkins, M. W. (2010b). Exploratory and higher-order factor analyses of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV) adolescent subsample. *School Psychology Quarterly*, *25*, 223–235.

Carretta, T. R., & Ree, J. J. (2001). Pitfalls of ability research. *International Journal of Selection and Assessment*, *9*, 325–335.

Cassidy, L. C, (1997). *The stability of WISC-III scores: For whom are triennial reevaluations necessary?* Unpublished doctoral dissertation, Kingston, RI: University of Rhode Island.

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

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research*, *30*, 429–452.

Carroll, J. B. (1997a). The three-stratum theory of cognitive abilities. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 183–208). New York: Guilford.

Carroll, J. B. (1997b). Theoretical and technical issues in identifying a factor of general intelligence. In B. Devlin, S. E. Fienberg, D. P. Resnick, & K. Roeder (Eds.), *Intelligence, genes, and success: Scientists respond to the bell curve* (pp. 125–156). New York: Springer-Verlag.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). New York: Pergamon Press.

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

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Cohen, J. (1959). The factorial structure of the WISC at ages 7–6, 10–6, and 13–6. *Journal of Consulting Psychology*, *23*, 285–299.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Boston: Addison–Wesley.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin*, *50*, 456–473.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington Publishers.

Daley, C. E., & Nagle, R. J. (1996). Relevance of WISC-III indicators for assessment of learning disabilities. *Journal of Psychoeducational Assessment*, *14*, 320–333.

Dana J., & Dawes, R. M. (2007). Comment on Fiorello et al., "Interpreting intelligence test results for children with disabilities: Is global intelligence relevant?" *Applied Neuropsychology*, *14*, 21–25.

Daniel, M. H. (2007). "Scatter" and the construct validity of FSIQ: Comment on Fiorello et al. (2007). *Applied Neuropsychology*, *14*, 291–295.

Daniel, M. H. (2009, August). *Subtest variability and the validity of WISC–IV composite scores*. Paper presented at the 2009 annual convention of the American Psychological Association, Toronto, ON, CA.

Davidow, J., & Levinson, E. M. (1993). Heuristic principles and cognitive bias in decision making: Implications for assessment in school psychology. *Psychology in the Schools*, *30*, 351–361.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: The Free Press.

Dawes, R. M. (2005). The ethical implications of Paul Meehl's work on comparing clinical versus actuarial prediction methods. *Journal of Clinical Psychology*, *61*, 1245–1255.

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

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence*. Boston: Allyn & Bacon.

DiStefano, C., & Dombrowski, S. C. (2006). Investigating the theoretical structure of the Stanford-Binet–Fifth Edition. *Journal of Psychoeducational Assessment*, *24*, 123–136.

Dombrowski, S. C., & Watkins, M. W. (in press). Exploratory and higher order factor analysis of the WJ-III full test battery: A school aged analysis. *Psychological Assessment.*

Dombrowski, S. C., Watkins, M. W., & Brogan, M. J. (2009). An exploratory investigation of the factor structure of the Reynolds Intellectual Assessment Scales (RIAS). *Journal of Psychoeducational Assessment*, *27*, 494–507.

Donders, J. (1996). Cluster subtypes in the WISC-III standardization sample: Analysis of factor index scores. *Psychological Assessment*, *8*, 312–318.

Dumont, R., Farr, L. P., Willis, J. O., & Whelley, P. (1998). 30-second interval performance on the coding subtest of the WISC-III: Further evidence of WISC folklore? *Psychology in the Schools*, *35*, 111–117.

Elliott, C. D. (1990). *Differential Ability Scales*. San Antonio, TX: The Psychological Corporation.

Elliott, C. D. (2007a). *Differential Ability Scales–2nd edition.* San Antonio, TX: The Psychological Corporation.

Elliott, C. D. (2007b). *Differential Ability Scales–2nd edition: Introductory and technical handbook.* San Antonio, TX: The Psychological Corporation.

Evans, J. J., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2001). The relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. *School Psychology Review*, *31*, 246–262.

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice*, *17*, 420–430.

Faust, D. (1990). Data integration in legal evaluations: Can clinicians deliver on their premises? *Behavioral Sciences and the Law*, *7*, 469–483.

Faust, D. (2007). Some global and specific thoughts about some global and specific issues. *Applied Neuropsychology*, *14*, 26–36.

Fiorello, C. A., Hale, J. B., Holdnack, J. A., Kavanagh, J. A., Terrell, J., & Long, L. (2007). Interpreting intelligence test results for children with disabilities: Is global intelligence relevant? *Applied Neuropsychology*, *14*, 2–12.

Fiorello, C. A., Hale, J. B., McGrath, M., Ryan, K., & Quinn, S. (2001). IQ interpretation for children with flat and variable test profiles. *Learning and Individual Differences*, *13*, 115–125.

Flanagan, D. P. (2000). Wechsler-based CHC cross-battery assessment and reading achievement: Strengthening the validity of interpretations drawn from Wechsler test scores. *School Psychology Quarterly*, *15*, 295–329.

Flanagan, D. P., Andrews, T. J., & Genshaft, J. L. (1997). The functional utility of intelligence tests with special education populations. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 457–483). New York: The Guilford Press.

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC–IV assessment.* Hoboken, NJ: Wiley.

Flanagan, D. P., & McGrew, K. S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 314–325). New York: Guilford.

Flanagan, D. P., & Ortiz, S. O. (2001). *Essentials of cross-battery assessment.* New York: Wiley.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2008). *Essentials of cross-battery assessment* (2nd ed.). New York: Wiley.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

Floyd, R. G., Keith, T. Z., Taub, G. E., & McGrew, K. S. (2007). Cattell-Horn-Carroll cognitive abilities and their effects on reading decoding skills: g has indirect effects, more specific abilities have direct effects. *School Psychology Quarterly*, *22*, 200–233.

Floyd, R. G., McGrew, K. S., & Evans, J. J. (2008). The relative contributions of the Cattell-Horn-Carroll cognitive abilities in explaining writing achievement during childhood and adolescence. *Psychology in the Schools*, *45*, 132–144.

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*, *35*, 169–182.

Freberg, M. E., Vandiver, B. J., Watkins, M. W., & Canivez, G. L. (2008). Significant factor score variability and the validity of the WISC-III Full Scale IQ in predicting later academic achievement. *Applied Neuropsychology*, *15*, 131–139.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice*, *4*, 99–120.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology*, *1*, 67–89.

Garb, H. N., & Boyle, P. A. (2003). Understanding why some clinicians use pseudoscientific methods: Findings from research on clinical judgment. In S. O. Lilienfeld, S. J. Lynn, & J. M. Lohr (Eds.), *Science and pseudoscience in clinical psychology* (pp. 17–38). New York: Guilford.

Glutting, J. J. (1986a). The McDermott Multidimensional Assessment of Children: Applications to the classification of

childhood exceptionality. *Journal of Learning Disabilities*, *19*, 331–335.

Glutting, J. J. (1986b). The McDermott Multidimensional Assessment of Children: Contribution to the development of individualized education programs. *Journal of Special Education*, *20*, 431–445.

Glutting, J. J., Adams, W., & Sheslow, D. (2000a). *Wide Range Intelligence Test*. Wilmington, DE: Wide Range, Inc.

Glutting, J. J., Adams, W., & Sheslow, D. (2000b). *Wide Range Intelligence Test: Manual.* Wilmington, DE: Wide Range, Inc.

Glutting, J. J., & McDermott, P. A. (1990a). Score structures and applications of core profile types in the McCarthy Scales standardization sample. *Journal of Special Education*, *24*, 212–233.

Glutting, J. J., & McDermott, P. A. (1990b). Patterns and prevalence of core profile types in the WPPSI standardization sample. *School Psychology Review*, *19*, 471–491.

Glutting, J. J., McDermott, P. A., & Konold, T. R. (1997). Ontology, structure, and diagnostic benefits of a normative subtest taxonomy from the WISC-III standardization sample. In D. P. Flanagan, J. L. Genshaft, & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 349–372). New York: Guilford.

Glutting, J. J., McDermott, P. A., Konold, T. R., Snelbaker, A. J., & Watkins, M. W. (1998). More ups and downs of subtest analysis: Criterion validity of the DAS with an unselected cohort. *School Psychology Review*, *27*, 599–612.

Glutting, J. J., McGrath, E. A., Kamphaus, R. W., & McDermott, P. A. (1992). Taxonomy and validity of subtest profiles on the Kaufman Assessment Battery for Children. *Journal of Special Education*, *26*, 85–115.

Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAI-II. *Journal of Special Education*, *40*, 103–114.

Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactored and cross-battery assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2nd ed., pp. 343–374). New York: Guilford.

Glutting, J. J., Youngstrom, E. A., Oakland, T., & Watkins, M. W. (1996). Situational specificity and generality of test behaviors for samples of normal and referred children. *School Psychology Review*, *25*, 94–107.

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. L. (1997). Incremental efficacy of WISC-III factor scores in predicting achievement: What do they tell us? *Psychological Assessment*, *9*, 295–301.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gottfredson, L. S. (1997). Intelligence and social policy. *Intelligence*, *24*, 288–320.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, *15*, 25–46.

Gottfredson, L. S. (2008). Of what value is intelligence? In A. Prifitera, D. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical assessment and intervention* (2nd ed., pp. 545–564). Amsterdam: Elsevier.

Gottfredson, L. S. (2009). Logical fallacies used to dismiss the evidence on intelligence testing. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 11–65). Washington, DC: American Psychological Association.

Groth-Marnat, G. (1997). *Handbook of psychological assessment* (3rd ed.). NY: Wiley.

Gridley, B. E., & Roid, G. H. (1998). The use of the WISC-III with achievement tests. In A. Prifitera & D. Saklofske (Eds.), *WISC-III clinical use and interpretation* (pp. 249–288). New York: Academic Press.

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

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

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw-Hill.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 407–434.

Gustafsson, J.-E., & Snow, R. E. (1997). Ability profiles. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 107–135). Westport, CT: Greenwood Press.

Hale, J. B., & Fiorello, C. A. (2001). Beyond the academic rhetoric of "g": Intelligence testing guidelines for practitioners. *The School Psychologist, 55,* 113–139.

Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner's handbook*. New York: Guilford.

Hale, J. B., Fiorello, C. A., Bertin, M., & Sherman, R. (2003). Predicting math competency through neuropsychological interpretation of WISC-III variance components. *Journal of Psychoeducational Assessment*, *21*, 358–380.

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Holdnack, J. A., & Aloe, A. M. (2007). Is the demise of IQ interpretation justified? A response to special issue authors. *Applied Neuropsychology*, *14*, 37–51.

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Hoeppner, J. B., & Gaither, R. A. (2001). WISC-II predictors of academic achievement for children with learning disabilities: Are global and factor scores comparable? *School Psychology Quarterly*, *16*, 31–55.

Hale, R. L. (1981). Cluster analysis in school psychology: An example. *Journal of School Psychology*, *19*, 51–56.

Hale, R. L., & Raymond, M. R. (1981). Wechsler Intelligence Scale for Children–Revised patterns of strengths and weaknesses as predictors of the intelligence achievement relationship. *Diagnostique*, *7*, 35–42.

Hale, R. L., & Saxe, J. E. (1983). Profile analysis of the Wechsler Intelligence Scale for Children–Revised. *Journal of Psychoeducational Assessment*, *1*, 155–162.

Hanna, G. S., Bradley, F. O., & Holen, M. C. (1981). Estimating major sources of measurement error in individual intelligence scales: Taking our heads out of the sand. *Journal of School Psychology*, *19*, 370–376.

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment*, *15*, 456–466.

Hildebrand, D. K., & Ledbetter, M. F. (2001). Assessing children's intelligence and memory: The Wechsler Intelligence Scale for Children–Third Edition and the Children's Memory Scale. In J. J. W. Andrews, D. H. Saklofske, & H. L. Janzen (Eds.), *Handbook of psychoeducational assessment: Ability,*

achievement, and behavior in children (pp. 13–32). New York: Academic Press.

Hills, J. R. (1981). *Measurement and evaluation in the classroom* (2nd ed.). Columbus, OH: Merrill.

Holland, A. M., & McDermott, P. A. (1996). Discovering core profile types in the school-age standardization sample of the Differential Ability Scales. *Journal of Psychoeducational Assessment*, *14*, 131–146.

Horn, J. L. (1988). Thinking about human abilities. In R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 645–685). New York: Plenum Press.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock, *WJ-R technical manual* (pp. 197–232). Chicago: Riverside.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*, 253–270.

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: Guilford.

Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. *Psychological Assessment*, *15*, 443–445.

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

Jencks, C., Bartlett, S., Corcoran, M., Crouse, J., Eaglesfield, D., Jackson, G., et al. (1979). *Who gets ahead? The determinants of economic success in America*. New York: Basic Books.

Jensen, A. R. (1992). Commentary: Vehicles of *g*. *Psychological Science*, *3*, 275–278.

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

Jones, W. T. (1952). *A history of Western philosophy.* New York: Harcourt, Brace.

Kahana, S. Y., Youngstrom, E. A., & Glutting, J. J. (2002). Factor and subtest discrepancies on the Differential Abilities Scale: Examining prevalence and validity in predicting academic achievement. *Assessment*, *9*, 82–93.

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

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2005). A history of intelligence test interpretation. In D. P. Flanagan and P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 23–38). New York: Guilford.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York: Wiley-Interscience.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York: Wiley.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

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. (2004a). *Kaufman Assessment Battery for Children–Second Edition.* Circle Pines, MN: AGS Publishing.

Kaufman, A. S., & Kaufman, N. L. (2004b). *Kaufman Brief Intelligence Test–Second Edition.* Circle Pines, MN: AGS Publishing.

Kaufman, A. S., & Lichtenberger, E. O. (2000). *Essentials of WISC-III and WPPSI-R assessment.* New York: Wiley.

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). Hoboken, NJ: Wiley.

Kavale, K. A., & Forness, S. R. (1984). A meta-analysis of the validity of Wechsler Scale profiles and recategorizations: Patterns or parodies? *Learning Disability Quarterly*, *7*, 136–156.

Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. *School Psychology Quarterly*, *14*, 239–262.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1992). Relative usefulness of elevation, variability, and shape information from WISC-R, K-ABC, and Fourth Edition Stanford Binet profiles in predicting achievement. *Psychological Assessment*, *4*, 426–432.

Konold, T. R., & Canivez, G. L. (2010). Differential relationships among WISC-IV and WIAT-II scales: An evaluation of potentially moderating child demographics. *Educational and Psychological Measurement*, *70*, 613–627.

Konold, T. R., Glutting, J. J., McDermott, P. A., Kush, J. C., & Watkins, M. W. (1999). Structure and diagnostic benefits of a normative subtest taxonomy developed from the WISC-III standardization sample. *Journal of School Psychology*, *37*, 29–48.

Kotz, K. M., Watkins, M. W., & McDermott, P. A. (2008). Validity of the General Conceptual Ability score from the Differential Ability Scales as a function of significant and rare interfactor variability. *School Psychology Review*, *37*, 261–278.

Krohn, E. J., & Lamp, R. E. (1999). Stability of the SB:FE and K-ABC for young children from low-income families: A 5-year longitudinal study. *Journal of School Psychology*, *37*, 315–332.

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

Kuusinen, J., & Leskinen, E. (1988). Latent structure analysis of longitudinal data on relations between intellectual abilities and school achievements. *Multivariate Behavioral Research*, *23*, 103–118.

Leonard, T., & Hsu, J. S. J. (1999). *Bayesian methods: An analysis for statisticians and interdisciplinary researchers*. Cambridge, UK: Cambridge University Press.

Levine, A. J., & Marks, L. (1928). *Testing intelligence and achievement.* New York: Macmillan.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching.* Englewood Cliffs, NJ: Prentice-Hall.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2006). Why questionable psychological tests remain popular. *The Scientific Review of Alternative Medicine*, *10*, 6–15.

Lipsitz, J. D., Dworkin, R. H., & Erlenmeyer-Kimling, L. (1993). Wechsler comprehension and picture arrangement

subtests and social adjustment. *Psychological Assessment*, *5*, 430–437.

Livingston, R. B., Jennings, E., Reynolds, C. R., & Gray, R. M. (2003). Multivariate analyses of the profile stability of intelligence tests: High for IQs, low to very low for subtest analyses. *Archives of Clinical Neuropsychology*, *18*, 487–507.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points." *Annual Review of Psychology*, *51*, 405–444.

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 3, pp. 1–59). Palo Alto, CA: Consulting Psychology Press.

Lubinski, D., & Humphreys, L. G. (1997). Incorporating general intelligence into epidemiology and the social sciences. *Intelligence*, *24*, 159–201.

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "intelligent testing" approach to the WISC-III. *School Psychology Quarterly*, *12*, 197–234.

Maller, S. J., & McDermott, P. A. (1997). WAIS-R profile analysis for college students with learning disabilities. *School Psychology Review*, *26*, 575–585.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Oxford, UK: Williams & Wilkins.

Matarazzo, J. D., & Herman, D. O. (1984). Base rate data for the WAIS-R: Test-retest stability and VIQ-PIQ differences. *Journal of Clinical Neuropsychology*, *6*, 351–366.

Mayes, S. D., Calhoun, S. L., & Crowell, E. W. (1998). WISC-III profiles for children with and without learning disabilities. *Psychology in the Schools*, *35*, 309–316.

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. San Antonio, TX: Psychological Corporation.

McClain, A. L. (1996). Hierarchical analytic methods that yield different perspectives on dynamics: Aids to interpretation. *Advances in Social Science Methodology*, *4*, 229–240.

McDermott, P. A. (1980). A computerized system for the classification of developmental, learning, and adjustment disorders in school children. *Educational and Psychological Measurement*, *40*, 761–768.

McDermott, P. A. (1981). Sources of error in psychoeducational diagnosis of children. *Journal of School Psychology*, *19*, 31–44.

McDermott, P. A. (1993). National standardization of uniform multisituational measures of child and adolescent behavior pathology. *Psychological Assessment*, *5*, 413–424.

McDermott, P. A. (1994). *National profiles in youth psychopathology: Manual of adjustment scales for children and adolescents.* Philadelphia, PA: Edumetric and Clinical Science.

McDermott, P. A. (1998). MEG: Megacluster analytic strategy for multistage hierarchical grouping with relocations and replications. *Educational and Psychological Measurement*, *58*, 677–686.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment*, *8*, 290–302.

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *The Journal of Special Education*, *25*, 504–526.

McDermott, P. A., & Glutting, J. J. (1997). Informing stylistic learning behavior, disposition, and achievement through ability subtests—Or, more illusions of meaning? *School Psychology Review*, *26*, 163–176.

McDermott, P. A., Glutting, J. J., Jones, J. N., & Noonan, J. V. (1989). Typology and prevailing composition of core profile types in the WAIS-R standardization sample. *Psychological Assessment*, *1*, 118–125.

McDermott, P. A., Glutting, J. J., Jones, J. N., Watkins, M. W., & Kush, J. C. (1989). Identification and membership of core profile types in the WISC-R national standardization sample. *Psychological Assessment*, *1*, 292–299.

McDermott, P. A., Goldberg, M. M., Watkins, M. W., Stanley, J. L., & Glutting, J. J. (2006). A nationwide epidemiological modeling study of learning disabilities: Risk, protection, and unintended impact. *Journal of Learning Disabilities*, *39*, 230–251.

McDermott, P. A., Green, L. F., Francis, J. M., & Stott, D. H. (1999). *Learning Behaviors Scale*. Philadelphia, PA: Edumetric and Clinical Science.

McDermott, P. A., & Hale, R. L. (1982). Validation of a systems-actuarial computer process for multidimensional classification of child psychopathology. *Journal of Clinical Psychology*, *38*, 477–486.

McDermott, P. A., Marston, N. C., & Stott, D. H. (1993). *Adjustment Scales for Children and Adolescents*. Philadelphia, PA: Edumetric and Clinical Science.

McDermott, P. A., & Watkins, M. W. (1985). *Microcomputer systems manual for McDermott Multidimensional Assessment of Children.* New York: The Psychological Corporation.

McDermott, P. A., & Weiss, R. V. (1995). A normative typology of healthy, subclinical, and clinical behavior styles among American children and adolescents. *Psychological Assessment*, *7*, 162–170.

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

McFall, R. M. (2000). Elaborate reflections on a simple manifesto. *Applied and Preventive Psychology*, *9*, 5–21.

McFall, R. M. (2005). Theory and utility—key themes in evidence-based assessment: Comment on the special section. *Psychological Assessment*, *17*, 312–323.

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: Guilford.

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

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

McGrew, K. S., Keith, T. Z., Flanagan, D. P., & Vanderwood, M. (1997). Beyond g: The impact of Gf-Gc specific cognitive ability research on the future use and interpretation of intelligence tests in the schools. *School Psychology Review*, *26*, 189–201.

McGrew, K. S., & Knopik, S. N. (1996). The relationship between intra-cognitive scatter on the Woodcock-Johnson Psycho-Educational Battery–Revised and school achievement. *Journal of School Psychology*, *34*, 351–364.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual. Woodcock-Johnson III.* Itasca, IL: Riverside Publishing.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, *4*, 268–273.

Meehl, P. E. (1959). Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology*, *13*, 102–128.

Meehl, P. E. (1979). A funny thing happened to us on the way to latent entities. *Journal of Personality Assessment*, *43*, 564–581.

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

Meehl, P. E. (2001). Comorbidity and taxometrics. *Clinical Psychology: Science and Practice*, *8*, 507–519.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–216.

Milligan, G. W., & Hirtle, S. C. (2003). Clustering and classification methods. In J. A. Schinka, & W. F. Velicer, J. A. (Eds.), *Handbook of psychology: Research methods in psychology*, Vol. 2 (pp. 165–186). Hoboken, NJ: Wiley.

Mueller, H. H., Dennis, S. S., & Short, R. H. (1986). A meta-exploration of WISC-R factor score profiles as a function of diagnosis and intellectual level. *Canadian Journal of School Psychology*, *2*, 21–43.

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

Naglieri, J. A. (1997). Planning, attention, simultaneous, and successive theory and the Cognitive Assessment System: A new theory-based measure of intelligence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 247–267). New York: Guilford.

Naglieri, J. A. (2003a). *Naglieri nonverbal ability test–Individual administration*. San Antonio, TX: Harcourt Assessment.

Naglieri, J. A. (2003b). Naglieri nonverbal ability tests: NNAT and MAT-EF. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 175–190). New York: Kluwer.

Naglieri, J. A., & Bornstein, B. T. (2003). Intelligence and achievement: Just how correlated are they? *Journal of Psychoeducational Assessment*, *21*, 244–260.

Naglieri, J. A., & Das, J. P. (1990). Planning, attention, simultaneous, and successive (PASS) cognitive processes as a model for intelligence. *Journal of Psychoeducational Assessment*, *8*, 303–337.

Naglieri, J. A., & Das, J. P. (1997a). *Cognitive Assessment System.* Itasca, IL: Riverside Publishing.

Naglieri, J. A., & Das, J. P. (1997b). *Cognitive Assessment System: Interpretive handbook.* Itasca, IL: Riverside Publishing.

National Association of School Psychologists. (2010). *Principles of professional ethics*. Bethesda, MD: NASP.

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

Nelson, J. M., & Canivez, G. L. (2012). Examination of the structural, convergent, and incremental validity of the Reynolds Intellectual Assessment Scales (RIAS) with a clinical sample. *Psychological Assessment*, *24*, 129–140.

Nelson, J. M., Canivez, G. L, Lindstrom, W., & Hatt, C. (2007). Higher-order exploratory factor analysis of the Reynolds Intellectual Assessment Scales with a referred sample. *Journal of School Psychology*, *45*, 439–456.

Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Erlbaum.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *12*, 248–277.

Nunnally, J. D., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Oh, H. J., Glutting, J. J., Watkins, M. W., Youngstrom, E. A., & McDermott, P. A. (2004). Correct interpretation of latent versus observed abilities: Implications from structural equation modeling applied to the WISC-III and WIAT linking sample. *Journal of Special Education*, *38*, 159–173.

Osgood, C. E., & Suci, G. J. (1952). A measure of relation determined by both mean differences and profile information. *Psychological Bulletin*, *49*, 251–262.

Piedmont, R. L., Sokolove, R. L., & Fleming, M. Z. (1989). An examination of some diagnostic strategies involving the Wechsler intelligence scales. *Psychological Assessment*, *1*, 181–185.

Pfeiffer, S. I., Reddy, L. A., Kletzel, J. E., Schmelzer, E. R., & Boyer, L. M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly*, *15*, 376–385.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills*, *105*, 997–1014.

The Psychological Corporation (1999). *Wechsler Abbreviated Scale of Intelligence.* San Antonio, TX: Psychological Corporation.

Public Law (P.L.) 108–446. Individuals with Disabilities Education Improvement Act of 2004 (IDEIA). (20 U.S.C. 1400 et seq.). 34 CFR Parts 300 and 301. Assistance to States for the education of children with disabilities and preschool grants for children with disabilities; Final Rule. *Federal Register*, *71* (156), 46540–46845.

Rapaport, D., Gil, M. M., & Schafer, R. (1945–1946). *Diagnostic psychological testing* (2 vols.). Chicago: Year Book Medical.

Ree, M. J., & Carretta, T. R. (1997). What makes an aptitude test valid? In R. F. Dillon (Ed.), *Handbook on testing* (pp. 65–81). Westport, CT: Greenwood Press.

Ree, M. J., Carretta, T. R., & Green, M. T. (2003). The ubiquitous role of *g* in training. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 262–274). New York: Pergamon Press.

Ree, M. J., Earles, J. A., & Treachout, M. S. (1994). Predicting job performance: Not much more than *g*. *The Journal of Applied Psychology*, *79*, 518–524.

Reinecke, M. A. Beebe, D. W., & Stein, M. A. (1999). The third factor of the WISC-III: It's (probably) not freedom from distractibility. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*, 322–328.

Reynolds, C. R. (1984). Critical measurement issues in assessment of learning disabilities. *Journal of Special Education*, *18*, 451–476.

Reynolds, C. R. (2007). Subtest level profile analysis of intelligence tests: Editor's remarks and introduction. *Applied Neuropsychology*, *14*, 1.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales*. Lutz, FL: Psychological Assessment Resources.

Riccio, C. A., Cohen, M. J., Hall, J., & Ross, C. M. (1997). The third and fourth factors of the WISC-III: What they don't measure. *Journal of Psychological Assessment*, *15*, 27–39.

Rispens, J., Swaab, H., van den Oord, E. J. C. G., Cohen-Kettenis, P., van Engeland, H., & van Yperen, T. (1997). WISC profiles in child psychiatric diagnosis: Sense or nonsense? *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*, 1587–1594.

Roid, G. H. (2003a). *Stanford-Binet Intelligence Scales–Fifth Edition*. Itasca, IL: Riverside Publishing.

Roid, G. H. (2003b). *Stanford-Binet Intelligence Scales–Fifth Edition: Technical manual*. Itasca, IL: Riverside Publishing.

Ryan, J. J., Kreiner, D. S., & Burton, D. B. (2002). Does high scatter affect the predictive validity of WAIS-III IQs? *Applied Neuropsychology*, *9*, 173–178.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology*, *56*, 573–605.

Salvia, J., & Ysseldyke, J. E. (1988). *Assessment in special and remedial education* (4th ed.). Boston: Houghton Mifflin.

Salvia, J., & Ysseldyke, J. E. (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin.

Saklofske, D. H. (2008). Forward. In D. Wechsler (Author), *Wechsler Adult Intelligence Scale–Fourth Edition.* San Antonio, TX: Pearson.

Sattler, J. M. (1982). *Assessment of children's intelligence and special abilities* (2nd ed.). Boston: Allyn & Bacon.

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Sattler, J. M. (1992). *Assessment of children* (3rd ed., revised and updated). San Diego, CA: Author.

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

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

Sattler, J. M. & Ryan, J. J. (2009). *Assessment with the WAIS-IV*. San Diego, CA: Author.

Schinka, J. A., & Vanderploeg, R. D. (1997). Profile clusters in the WAIS-R standardization sample. *Journal of the International Neuropsychological Society*, *3*, 120–127.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274.

Schneider, J. (2008). Playing statistical Ouija board with commonality analysis (and other errors). *Applied Neuropsychology*, *15*, 44–53.

Smith, C. B., & Watkins, M. W. (2004). Diagnostic utility of the Bannatyne WISC-III pattern. *Learning Disabilities Research and Practice*, *19*, 49–56.

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

Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

Tatsuoka, M. M. (1974). *Classification procedures: Profile similarity.* Champaign, IL: Institute for Personality and Ability Testing.

Tatsuoka, M. M., & Lohnes, P. R. (1988). *Multivariate analysis* (2nd ed.). New York: Macmillan.

Taub, G. E., Keith, T. Z., Floyd, R. G., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly*, *23*, 187–198.

Teeter, P. A., & Korducki, R. (1998). Assessment of emotionally disturbed children with the WISC-III. In A. Prifitera & D. H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 119–138). New York: Academic Press.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior*, *29*, 332–339.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Technical manual, Stanford-Binet Intelligence Scale: 4th edition.* Chicago, IL: Riverside.

Thorndike, R. M. (1990). Origins of intelligence and its measurement. *Journal of Psychoeducational Assessment*, *8*, 223–230.

Thorndike, R. M. (1997). The early history of intelligence testing. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues.* New York: Guilford.

Tulsky, D. S., Saklofske, D. H., Chelune, G. J., Heaton, R. K., Ivnik, R. J., Bornstein, R., et al. (2003). *Clinical interpretation of the WAIS-III and WMS-III*. San Diego, CA: Academic Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Ward, J. J., Jr. (1963). Hierarchical grouping to optimize an objective function. *American Statistical Association Journal*, *58*, 236–244.

Ward, S. B., Ward, T. B., Hatt, C. V., Young, D. L., & Mollner, N. R. (1995). The incidence and utility of the ACID, SCIDS, and SCAD profiles in a referred population. *Psychology in the Schools*, *12*, 267–276.

Watkins, M. W. (1996). Diagnostic utility of the WISC-III developmental index as a predictor of learning disabilities. *Journal of Learning Disabilities*, *29*, 305–312.

Watkins, M. W. (1999). Diagnostic utility of WISC-III subtest variability among students with learning disabilities. *Canadian Journal of School Psychology*, *15*, 11–20.

Watkins, M. W. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly*, *15*, 465–479.

Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion? *The Scientific Review of Mental Health Practice*, *2*, 118–141.

Watkins, M. W. (2005). Diagnostic validity of Wechsler subtest scatter. *Learning Disabilities: A Contemporary Journal*, *3*, 20–29.

Watkins, M. W. (2006). Orthogonal higher-order structure of the WISC-IV. *Psychological Assessment*,*18*,123–125.

Watkins, M. W. (2009). Errors in diagnostic decision making and clinical judgment. In T. B. Gutkin & C. R. Reynolds (Eds.), *Handbook of school psychology* (4th ed., pp. 210–229). Hoboken, NJ: Wiley.

Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite strengths and weaknesses. *Psychological Assessment*, *16*, 133–138.

Watkins, M. W., Lei, P., & Canivez, G. L. (2007). Psychometric intelligence and achievement: A cross-lagged panel analysis. *Intelligence*, *35*, 59–68.

Watkins, M. W., & Glutting, J. J. (2000). Incremental validity of the WISC-III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment*, *12*, 402–408.

Watkins, M. W., Glutting, J. J., & Lei, P.-W. (2007). Validity of the full-scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology*, *14*, 13–20.

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (2005). Issues in subtest profile analysis. In D. P. Flanagan and P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed, pp. 251–268). New York: Guilford.

Watkins, M. W., & Kush, J. C. (1994). Wechsler subtest analysis: The right way, the wrong way, or no way? *School Psychology Review*, *23*, 638–649.

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997a). Discriminant and predictive validity of the WISC-III ACID profile among children with learning disabilities. *Psychology in the Schools*, *34*, 309–319.

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997b). Prevalence and diagnostic utility of the WISC-III SCAD profile among children with disabilities. *School Psychology Quarterly*, *12*, 235–248.

Watkins, M. W., Kush, J. C., & Schaefer, B. A. (2002). Diagnostic utility of the learning disability index. *Journal of Learning Disabilities*, *35*, 98–103.

Watkins, M. W., Wilson, S. M., Kotz, K. M., Carbone, M. C., & Babula, T. (2006). Factor structure of the Wechsler Intelligence Scale for Children–Fourth Edition among referred students. *Educational and Psychological Measurement*, *66*, 975–983.

Watkins, M. W., & Worrell, F. C. (2000). Diagnostic utility of the number of WISC-III subtests deviating from mean performance among students with learning disabilities. *Psychology in the Schools*, *37*, 303–309.

Wechsler, E. (1939). *The measurement of adult intelligence.* Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children.* New York: Psychological Corporation.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children–Revised.* New York: Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–3rd Edition.* San Antonio, TX: Psychological Corporation.

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

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–3rd Edition.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1999). *The Wechsler Intelligence Scale for Children—3rd Edition (Swedish version).* Stockholm, Sweden: Psykologiförlaget.

Wechsler, D. (2002a). *WAIS-III/WMS-III technical manual, updated.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2002b). *Wechsler Preschool and Primary Scale of Intelligence–3rd Edition.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children–4th Edition.* San Antonio, TX: Psychological Corporation.

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

Wechsler, D. (2008b). *Wechsler Adult Intelligence Scale–4th Edition: Technical and interpretive manual.* San Antonio, TX: Pearson.

Wechsler, D., & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability.* San Antonio, TX: The Psychological Corporation.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment*, *53*, 827–831.

Weiss, L. G., & Prifitera, A. (1995). An evaluation of differential prediction of WIAT achievement scores from WISC-III FSIQ across ethnic and gender groups. *Journal of School Psychology*, *33*, 297–304.

Weiss, L. G., Saklofski, D. H., & Prifitera, A. (2003). Clinical interpretation of the Wechsler Intelligence Scale for Children–Third Edition (WISC-III) Index scores. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2nd ed., pp. 115–146). New York: Guilford Press.

Wiggins, J. S. (1988). *Personality and prediction: Principles of personality assessment.* Malabar, FL: Krieger Publishing Company.

Wilhoit, B. E., & McCallum, R. S. (2002). Profile analysis of the Universal Nonverbal Intelligence Test standardization sample. *School Psychology Review*, *31*, 263–281.

Wolber, G. J., & Carne, W. F. (2002). *Writing psychological reports: A guide for clinicians* (2nd ed.). Sarasota, FL: Professional Resources Press.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson–Revised Tests of Achievement*. Itasca, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities.* Itasca, IL: Riverside Publishing.

Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests.* New York: Henry Holt & Company.

Youngstrom, E. A., Kogos, J. L., & Glutting, J. J. (1999). Incremental efficacy of Differential Ability Scales factor scores in predicting individual achievement criteria. *School Psychology Quarterly*, *14*, 26–39.

Zachary, R. A. (1990). Wechsler's intelligence scales: Theoretical and practical considerations. *Journal of Psychoeducational Assessment*, *8*, 276–289.

Zander, E., & Dahlgren, S. O. (2010). WISC-III Index Score profiles of 520 Swedish children with pervasive developmental disorders. *Psychological Assessment*, *22*, 213–222.

Zeidner, M. (2001). Invited foreword and introduction. In J. J. W. Andrews, D. H. Saklofske, & H. L. Janzen (Eds.), *Handbook of psychoeducational assessment: Ability, achievement, and behavior in children*. New York: Academic Press.

Zhu, J., & Weiss, L. (2005). The Wechsler scales. In D. P. Flanagan and P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 297–324). New York: Guilford.