# Three Practice Parameters for Interpreting Intelligence Test Part Scores Psychological Asessment Research Team University of Georgia

Heidi Zinzow, Meghan VanDeventer, Kelley Tarwater, Carolyn Stephens, Chandler Sims, Devlynne Shannon, Puja Seth, Megan Benoit, Nicholas Polizzi, Justin Miller, Carrah Martin, Kyongboon Kwon, R. W. Kamphaus, Tamerah Hunt, Jennifer Hartwig, Elizabeth Haman, Mauricio A. Garcia-Barrera, and Anny Castilla.

## Resumen

El mundo de las pruebas de inteligencia en niños ha tenido cambios dramáticos durante las últimas dos décadas debido a la publicación de muchas pruebas nuevas, acompañadas de nuevos puntajes derivados o parciales. Estos puntajes parciales se utilizan algunas veces como aproximaciones a las medidas de inteligencia general (es decir, los puntajes totales compuestos). Dado el surgimiento de una diversidad de puntajes parciales, los psicólogos y los estudiantes de psicología tratan de clarificar en que circunstancias se deben utilizar estos puntajes parciales. Basados en la evidencia existente, proponemos se adopten parámetros prácticos para asegurarnos que se informe la estimación de la inteligencia general y su relación con los puntajes parciales. Se proponen tres parámetros prácticos. Primero, en prácticamente todos los casos, en la evaluación de la inteligencia general y en la toma de decisiones diagnosticas, se debería utilizar el puntaje de inteligencia compuesto. Segundo, en circunstancias especiales, las medidas de habilidad verbal/cristalizada o habilidad de razonamiento/fluida son los mejores puntajes parciales a utilizar en vez del puntaje compuesto general. Tercero, los puntajes visual/espacial/no verbal, memoria de trabajo/a corto término, velocidad de procesamiento, y similares, pueden utilizarse con propósitos de tamizaje, pero no para tomar decisiones diagnosticas. Se presentan algunas conclusiones relativas a los componentes de la inteligencia analizados a través de sus puntajes.

*Palabras claves:* inteligencia, puntajes compuestos, puntajes parciales.

## Summary

The practical world of children's intelligence testing has undergone convulsive change in the last two decades due to the publication of many new tests, accompanied by new derived or "part" scores. These part scores are sometimes used as proxies for measures of general intelligence (i.e. overall composite scores). Because of the emergence of new and a larger variety of part scores, psychologists and psychologists in training are seeking clarification regarding the circumstances under which part scores should be used in lieu of overall intelligence scores. Based upon abundant empirical evidence, we propose that practice parameters be adopted to ensure that estimation of general intelligence is informed by the wealth of scientific knowledge on the assessment of general intelligence and the relationship of part scores to this process. Three practice parameters are offered. First, in virtually all cases, the overall composite intelligence test score should be used for the assessment of general intelligence and diagnostic decision making. Second, under extraordinary and in rare circumstances measures of verbal/crystallized ability or reasoning/fluid ability are the best part scores to use in lieu of the overall composite score. Third, visual/spatial/nonverbal, working/short-term memory, processing speed, and like scores may be used for screening purposes but not for making diagnostic decisions**.** Conclusions regarding the compartmentalization of intelligence via part scores are offered.

*Keywords:* intelligence measures, compound scores, part scores.

## Three Practice Parameters for Interpreting Intelligence Test Part Scores

The practice of children's intelligence testing has undergone unprecedented change in the last two decades. The overwhelming evidence for the Flynn effect that causes test score norms to change with rhythmical regularity, Carroll's (1993) Herculean re-analysis of over 450 data sets, Hart and Risley's (1995) patient and thorough documentation of the origins of verbal ability in infancy, and the publication of new Binet V (Roid, 2003), WISC-IV (Wechsler, 2003), the Reynolds Intellectual Assessment Scale ([RIAS], Reynolds & Kamphaus, 2003) and other scales in 2003 alone, call psychologists to change their intellectual assessment practices. Even the venerable WISC has changed significantly in that the 2003 edition abandoned the Verbal and Performance IQs and replaced them with four part scores. The emergence of a variety of part scores on modern intelligence tests requires practicing psychologists to develop a thorough understanding of the validity of these new scores.

Knowledge of part score validity is especially important for the assessment of general intelligence. Although the aforementioned sounds incongruent it is nevertheless true. Under numerous circumstances such as bilingualism, cultural differences, regional differences, lack of assessment time, screening, and disdain for the concept of general intelligence, psychologists may use part scores to estimate a level of general intelligence. The part score may be used for making important diagnostic decisions as well, such as making a diagnosis of mental retardation, screening for giftedness, or identifying a "normal" level of intelligence for assessing a learning disability. These uses of part scores are particularly frequent for the assessment of children, the most prevalent use of intelligence tests world wide. Luckily, there is a substantial body of literature on the relationship of overall composite to part scores for the assessment of children's intelligence.

The wealth of available intelligence assessment research and cogent reviews of the literature make it possible to draw practice parameters that transcend the specific changes in scales and tests that have occurred (see Sternberg, Grigorenko, & Bundy, 2001, for an extensive review of the predictive validity of the overall intelligence test composite score; Lubinski, 2004, for a review of the scientific evidence in support of the hierarchical organization of intelligence consistent with general intelligence theory; Kanaya, Scullin, & Ceci, 2003, for a summary of research on the Flynn Effect; Watkins & Glutting, 2000, and Watkins & Canivez, 2004, for a review of the literature on subtest profile analysis; and Moffitt, Caspi, Harkness, & Silva, 1993, for a documentation of the literature supporting the stability of overall composite intelligence test scores over development, among many other extensive reviews of the literature). Lubinski (2004) commented on the size and consistency of intelligence scientific literature by observing that;

> The study of individual differences in cognitive abilities is one of the few branches of psychological science to amass a coherent body of empirical knowledge withstanding the test of time. (p. 96)

This article represents an attempt to begin to fill a gap in psychology practice literature regarding the use of part versus overall composite scores for the assessment of children's intelligence, a void that has been created by the proliferation of part scores on the Wechsler, Binet, and other well known intelligence tests. The stakes for patients are high.

### ***Practice Parameter 1***

*In virtually all cases, the overall composite intelligence test score should be used for diagnostic decision making.* This proposition is made based on strong and consistent evidence that the overall composite score offered by intelligence tests is the most stable of all scores in an intelligence test battery, has the most predictive validity, and is supported by strong scientific and theoretical support.

#### *Stability of Overall Composite*

The stability of IQ scores has been a topic of discussion ever since Sterm established the concept of an intelligence quotient (IQ) in 1912 (Hopkins & Bibelheimer, 1971). A great deal of research has indicated that intelligence test scores are fairly stable among individuals (Canivez & Watkins, 2001; Neisser, et al., 1996; Raguet, Campbell, Berry, Schmitt, & Smith, 1996). Estimates of the stability coefficients have varied across studies; however, Bloom (1964) concluded that intelligence for a 13-year-old predicts at least 90% of the variance in intelligence for 18-year olds. Gustaffson and Olav-Undeheim (1992) found that for general intelligence, scores at age 12 predicted 85% of the variance in intelligence for 15-year-olds, which is somewhat lower than some previous findings but still similar.

Moffit et al. (1993) examined whether changes in intelligence are reliable and whether the reliable changes were systematic or random changes. They conducted a longitudinal study with children who were part of the Multidisciplinary Health and Development Study in Dunedin, New Zealand. Participants included 991 children at age five, 954 at age seven, 955 at age nine, 925 at age 11, and 850 at age 13. The participants were administered the Wechsler Intelligence Scale for Children-Revised (WISC-R) in order to obtain an IQ score for the children.

The results indicated that there were changes in IQ scores across time with 107 children who were placed in the labile group. When examining the cumulative and sequential changes in this group, it was by 5.3 IQ points. Therefore, the changes did not appear to be very meaningful. Even though for some children the changes in their IQ score surpassed that expected by measurement error, the changes could be explained by measurement error for most individuals. The results also indicated that the pattern of change in IQ scores appear to be unreliable changes; however, for the small group of children whose changes surpassed expectations, the profiles of IQ change were reliable, but they were not systematic. When examining family contextual factors and individual differences in children that might be correlated with IQ change, it was found that the stable and labile groups were very similar. Therefore, it is difficult to determine the characteristics of the individuals whose IQ scores change over a period of time (Moffitt et al., 1993).

Canivez and Watkins (1998) also examined the long-term stability of the WISC-III in 667 students who had been evaluated for special education. The results indicated that IQ scores and factor index scores with stability coefficients ranging from .62 to .91 were more stable than the subtest scores with stability coefficients ranging from .55 to .78. Overall, these results tended to support the long-term stability of IQ scores for children who took the WISC-III.

A more recent study conducted by Canivez and Watkins (2001) examined the stability of the WISC-III in 522 children with disabilities. The results indicated that there were no differences in the stability coefficients for IQ scores, index scores, and subtest scores between the specific learning disability, serious emotional disability, and mental retardation groups. In addition, it was found that the Full Scale IQ (FSIQ) score was adequately stable across all three groups, and stability coefficients for FSIQ ranged from .85-.90. However, the stability of the index scores and subtest scores were inadequate across this clinical sample. These results support previous findings (Canivez & Watkins, 1998) suggesting that overall global IQ scores tend to be fairly stable across time.

Although a great deal of research supports the stability of general intelligence, it is difficult to determine the reason for the lack of change. One possibility is the stability of environmental factors. Sameroff, Seiffer, Baldwin, and Baldwin (1993) found a stability coefficient of .72 when examining children at age four and then at age 13. However, risk factors, such as family social support, major stressful life events, disadvantaged minority status, mother’s mental health, and others, were also fairly high during this time with a coefficient of .76. The researchers concluded that the cumulative amount of risk present in the child’s environment was a better predictor of intelligence then pattern of risk. Therefore, genetics along with a “fixed” high-risk environment might contribute to the stability of intelligence scores.

*Instability of Score Differences and Profiles*

Several studies have questioned the usefulness and stability of profile and discrepancy scores in diagnosing learning disabilities. The question of score stability is important since the choice to use a certain score or profile might affect educational placement, qualification for government assistance programs, and even capital punishment decisions. An investigation by Canivez and Watkins (2001) demonstrated that Full-scale IQ scores are the most stable within and between major disability groups including children diagnosed with Learning Disabilities, Mental Retardation, and emotional disturbances. The stability ranges of subtests and discrepancy scores between Verbal IQ and Performance IQ were not acceptable (Canivez & Watkins, 2001).

A study by Shaywitz, Fletcher, Holahan, and, Shaywitz (1992) compared groups of children with reading disabilities from the Connecticut Longitudinal Study. The sample used for the study was a cohort of 445 children who started Kindergarten in 1983 at one of Connecticut's public schools. Within the sample, all the students were English speakers and none of them had significant sensory or psychiatric impairment. This group, assembled from a two-stage probability sample survey, was assessed periodically with academic measures, behavior scales, and intelligence tests (Shaywitz et al., 1992). From this sample, comparisons were made between a control group with no reading difficulties, a group diagnosed with Learning Disabilities using discrepancy scores between IQ and achievement tests, and a group diagnosed just by low reading achievement scores. This investigation found that there was no advantage to using the discrepancy model even though it is the most common method for identifying Learning Disabilities. Actually, the Full Scale IQ score, rather than the difference score, appeared to be the factor most significant for differentiating between the groups of low readers (Shaywitz et al., 1992). Additionally, the discrepancy model does not take into account the effects of regression, so that children with IQ's above the mean are more likely to have large difference scores and be over- identified for Learning Disabilities (Shaywitz et al., 1992). The follow- up study examined children from the Connecticut Longitudinal Study at adolescence and confirmed that the findings remained stable over time (Shaywitz et al. 1999). A similar investigation (Fletcher, Francis, Rourke, Shaywitz, & Shaywitz, 1992) comparing groups of children across four different methods of Learning Disability diagnosis, also showed little validity for the use of discrepancy scores.

D'Anguilli and Siegal (2003) used the Wechsler Intelligence Scale for Children-Revised (WISC-R) to compare three groups of children: a group with reading disabilities, a group with a specific arithmetic disability, and a group with typical achievement. Although significant differences between Verbal and Performance IQ scores were predicted to occur most frequently among the groups with learning disabilities, 65% or more of the subjects in these groups did not show the expected pattern (D'Anguilli & Siegal, 2003). The percentage of students in the typical achievement group who had large differences between the Verbal and Performance scales was not significantly smaller than the percentage of children with large differences in either of the learning disability groups. The patterns of scatter among the scores were statistically unimportant and not reliable enough to make predictions about groups or individual diagnosis (D'Anguilli & Siegal, 2003).

Profile analysis is the practice of interpreting patterns of subtest scores to assess individuals' strengths and weaknesses, which are thought to be useful in determining diagnoses and/or appropriate interventions (Glutting, McDermott, Konold, Snelbaker, & Watkins, 1998). Profile analyses can be of two ilk, normative and ipsative. Normative subtest profiles are those that compare an individual's subtest scores with those of a norm-referenced group (Kamphaus, 2001). Ipsative analyses are those that look at intraindividual differences, or those between the subtest scores of one person (Kamphaus, 2001). In particular, the Wechsler scales have been conducive to profile examination due to their provision of subtest standard scores. Clinical inferences have been made from the pattern of individuals' Wechsler subtest scores for more than five decades (Glutting et al., 1998). Indeed, clinical interpretations based on profile analyses are commonplace (Glutting et al., 1998; Livingston, Jennings, Reynolds, & Gray, 2003; McDermott & Glutting, 1997). Clinicians probably use subtest score profiles because they assume that the profiles are relatively stable (Livingston et al., 2003) and are providing unique and important information about their clients (McDermott & Glutting, 1997). Interestingly, the widespread use of profile analysis is not supported by empirical evidence.

Despite the apparent popularity of subtest score interpretation, several recent studies call its practice into question (Glutting et al., 1998; Glutting, McDermont, & Konold, 1997; Livingston et al., 2003; McDermott & Glutting, 1997; Watkins & Glutting, 2000). More pointedly, Watkins and Glutting (2000) found that the profile scatter in WISC-III scores did a very poor job of predicting achievement outcomes in both normal and exceptional students. They also found that, while profile shape accounted for a small portion of the variance in achievement, these results may have been inflated by measurement error, were intuitive (as in, those with low arithmetic subtest scores had lower math achievement), and were thus uninformative. In sum, Watkins and Glutting concluded that the incremental validity of scatter and shape profiles is of no predictive import.

Additional support for the argument made by Watkins and Glutting (2000) was garnered from several psychometric phenomena. First, as Livingston and colleagues (2003) pointed out, when ipsative profiles are calculated, the most dependable component of variance, general intelligence (g), is removed. This is a problem, according to Livingston et al., because the variance left behind is so unreliable as to be unworthy of interpretation. In addition to being less reliable than IQ and index score profiles, Livingston and colleagues pointed out that the instability of subtest score profiles makes them useless for clinical purposes. Second, there is a base rate problem in interpreting subtest profiles (Glutting, McDermott, Watkins, Kush, & Konold, 1997). That is, while people with various disabilities may exhibit unusual subtest profiles, they do so at rates no different than those of the general population. For example, Glutting, McDermott, Watkins, et al. (1997) found, “in essence, children with LD and ED were no more likely to exhibit exceptional subtest configurations than children in general” (p. 181). They asserted that, due to base rate misconceptions, a Barnum effect (after P.T. Barnum’s popular circuses, which offered something for everyone) exists in the profile analysis of ability scores. That is, profiles with high base rates (in other words, ordinary) somehow drew attention and interpretation from clinicians. Kaufman (1990) found that it was common for clinicians to underestimate how much scatter is normal among the population (as cited in Kamphaus, 2001). If attempting to infer diagnosis from a profile, this lack of knowledge about base rates can cause an overestimation of pathology (Kamphaus, 2001). Third, inverse probabilities and circularity of logic taint the use profile analysis (Glutting et al., 1998). According to Glutting and colleagues (1998), subtest profiles are used to form the selected group as well as to define it, the methodological problem of self-selection. Also, in terms of inverse probabilities, what is usually done to create a typical profile is testing a group of similarly diagnosed individuals and looking for a similarity in profiles (Glutting et al., 1998). However, this is the inverse of clinical reality, wherein patients are referred to the clinician in order that a diagnosis may be made or ruled out on the basis of a test (or the presence or absence of a particular profile) (Glutting et al., 1998). As Glutting and his colleagues (1998) pointed out, these situations are rarely equivalent and it may be poor science to treat them as such. Finally, there is simply a dearth of research in support of the validity of subtest profiles. As Kamphaus (2001) noted, the few studies that have been done (for example, Matheson, Mueller, & Short, 1984; Naglieri, Kamphaus, & Kaufman, 1983; as cited in Kamphaus, 2001) have met with unsupportive or inconclusive results. Kamphaus concluded that profile analysis does not have a sound research base and should not be used to infer causal relationships or to make diagnoses (of mental retardation, learning disabilities, or otherwise).

Macmann and Barnett (1997) raised several concerns with Kaufman’s *Intelligent Testing* approach and the use of profile analysis and ipsative patterns. They questioned the validity of hypotheses that are developed through score comparisons, specifically differences between Verbal IQ and Performance IQ, because they are inherently less reliable than composite scores. Since the scores involved in these comparisons are correlated, the resulting difference scores are less reliable than the scores that originally provided the basis for the comparison (Nunnally, 1978). Overall, research suggests that the Full-scale IQ scores seem to be very stable over-time, while the subtest score profiles are inherently less stable and provide little useful clinical information (Livingston et al., 2003).

Considering current research on the problems with subtest profile interpretation and its accompanying issues, it seems prudent to stop the widespread use of profile analyses until such time as sound scientific evidence for its use is found. It is the recommendation of this research team, and a position that echoes that of McDermott and Glutting (1997), that neither ipsative nor normative profile analysis be used when making diagnostic decisions.

Instead, clinicians should refer to the longstanding and well-researched validity and reliability of composite scores (Kamphaus, 2001; McDermott & Glutting, 1997), and make those the first and only choice when considering which IQs to use.

*Predictive Validity*

The overall composite intelligence test score is also the most useful for predicting the occupational success of adults. It is well known that very large individual differences exist among workers in productivity and job performance. However, it has been difficult to distinguish between citizenship behavior evaluation and performance evaluation. When conducting worker evaluations, supervisors tend to give more importance to the social behavior of the worker rather than to his or her productivity. This confusion has led to a belief that intelligence is not a determinant of job success. Researchers Hunter and Schmidt (1996) have contributed to the intelligence research field by focusing their work on the relationship between intelligence and job performance. Intelligence (general cognitive ability) has proven to be the main determinant of variation in job performance.

A study conducted by Hunter and Hunter (1984), based on performance ratings and training success measures, estimated the validity of intelligence for high-complexity jobs to be .57, for medium-complexity jobs .51, and for low-complexity jobs .38. Integrity, a personality trait, was found to be the next most valid predictor of job performance. The combination of intelligence and integrity yielded a validity of .65. Other important abilities that are significant to different jobs, such as psychomotor, social, and physical, have been shown to be less stable predictors. Intelligence has been found to be the predictor with the highest validity across different jobs; in other words, the validity of cognitive ability tests in employment is generalizable to different occupations.

Although some researchers suggest that experience is a more valid predictor of job performance, advocates of intelligence have shown the contrary. Schmidt, Hunter, Outerbridge, and Goff (1988) demonstrated that for a five-year period of experience the difference between job knowledge and supervisor performance ratings due to abilities remained constant over time. McDaniel, Schmidt, and Hunter (1988) found that the relationship between experience and job performance was not significant. They found a correlation of .49 between experience and performance correlation, which dropped to .25 with 6-9 years of experience, and to .15 when experience was 12 years or more. Hunter and Hunter (1984) found the average predictive validity for experience to be .18, whereas it is .51 for intelligence.

The theory supporting these findings consists of two main ideas. The first idea is based on Thorndike’s classic learning theory, which stated that the main determinant of individual differences in performance is the differences in learning among individuals. Due to the fact that intelligence predicts effectiveness of learning, it is hypothesized that intelligence would predict job performance. The second idea, supporting this prediction, is based on factor analytic studies of human performance with 30 specific cognitive skills. The composite skill test has been proven to be a measure of intelligence. Therefore, the general belief about the importance of specific cognitive skills implies that intelligence should predict job performance.

Hunter and Schmidt (1996) presented data from civilian and military populations that support these predictions.  They found the correlation between ability and knowledge to be .80 for the civilians and .63 for the military, between knowledge and performance to be .80 for civilians and .70 for the military, and between intelligence and performance to be .75 for civilians and .53 for the military. Thus, the theory for prediction is verified by the data.  Although social policy has limitations in recognizing the importance of using intelligence measures in hiring processes, there is no doubt that not taking intelligence into account is counterproductive and can lead to performance decrements (Hunter & Schmidt, 1996).

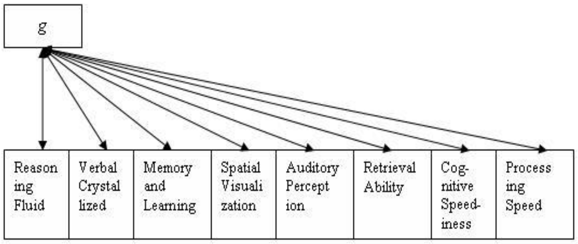
Another consistent and intriguing finding is that not only are overall intelligence test scores predictive of important child and adult outcomes, the accumulation of data for newer tests suggests that the "indifference of the predictor" phenomenon may be at work (Sternberg, Grigorenko, & Bundy, 2001). In other words, the choice of intelligence test is of little importance in that the composite score from all measures studied (e.g. Differential Ability Scales, K-ABC, Wechsler, Binet, etc.) demonstrate similar predictive validity results.

***Practice Parameter 2***

*Under extraordinary and in rare circumstances measures of verbal/crystallized ability or reasoning/fluid ability are the best part scores to use in lieu of the overall composite score.* We think that the evidence is strong that the only part scores that may be used in unusual circumstances for making diagnostic decisions are those that have been shown to measure either verbal/crystallized ability or reasoning/fluid ability. Our rationale is based on the consistent finding that measures of these latent constructs have always been virtually co-equal predictors of academic achievement in criterion-related and predictive validity studies (see Table 1).

Carroll's t*hree stratum theory* is a useful framework for understanding part scores. Carroll's (1993) hierarchical factor analyses yielded three "strata" of factors, the highest of which represents the familiar construct of general intelligence (*stratum three*) (Carroll, 1993).  In accordance with other supporters of general intelligence (Spearman, Cattell, Horn and others) Carroll believed that this construct accounts for much of the variance in intelligence test performance.  In this way, measures of the "narrow" first-strata traits are dominated by second-strata traits which, in turn, are dominated by stratum-three "g" (Kamphaus, 2001).  For example, much of the variance in a stratum-one factor such as spelling ability is accounted for by the other two strata.  Scores on a spelling test are determined first by general intelligence, second by the second-stratum factor of crystallized ability, and third by the specific spelling ability factor.

RIAS have been constructed with the intention of measuring some of these stratum two traits.  Therefore the second-stratum factors have numerous implications for the interpretation of intelligence tests. Carroll found that these factors varied in the strength of their relationship to "g" and their ability to predict important life outcomes. Those factors with the highest loadings on "g" are better predictors of life outcomes such as academic achievement and occupational success due to the fact that general intelligence is the best predictor of life outcomes.



We propose that second stratum factors other than reasoning/fluid ability and verbal/crystallized ability, such as working/ short-term memory and processing speed, should not be used as a substitute for a composite score when making a clinical diagnosis. When the validity of a composite score is questioned, only index scores that have adequate predictive validity and g-loadings can be used as a substitute for a composite score. Kaufman (1994) suggested a convention for rating the value of subtests g-loadings: g-loadings of .70 and above are considered “good”, .50 to .69 are “fair”, and g-loadings below .50 are “poor.” In a multiple-instrument factor analysis of the Woodcock-Johnson Psycho-Educational Battery and the Cognitive Assessment System, both crystallized (*Gc*) and the fluid (*Gf*) factors load at moderate to high levels (.60s -.70s) on the g-factor. Other factors such as long-term retrieval, *Glr*, short-term memory, *Gsm*, and auditory processing, *Ga*, load at much lower levels (.30s -.50s). Visual (*Gv*) and spatial (*Gs*) subtests tend to moderately load on g, ranging from .50s and .60s (Keith, T. Z., personal communication, June, 2001, in Committee on Disability Determination for Mental Retardation, 2002). In Carroll’s (1993) second hierarchical “broad” stratum, fluid intelligence and crystallized intelligence are most saturated with general intelligence. The correlation of the other factors with “g” decreases as they become distant from the general intelligence. The processing speed factor in the second stratum is the least correlated with general intelligence.

The inadequacy of second stratum factors such as working/short-term memory and processing speed as substitutes of general intelligence can be supported by the fact that those traits are not included in some well-established brief intelligence tests. Brief tests of intelligence such as the Wechsler Abbreviated Scale of Intelligence ([WASI], Psychological Corporation, 1999) and the Kaufman Brief Intelligence Test ([K-BIT], Kaufman & Kaufman, 1990) provide a valid measure of intelligence, yield crystallized and fluid abilities, and have strong psychometric properties. A four-subtest form of the WASI consists of Vocabulary, Similarities, Block Design, and Matrix Reasoning. The manual provides evidence for the validity of the WASI as a quick screening measure of general intellectual ability. The K-BIT consists of Vocabulary and Matrices and is reported to have high g-loadings (Kaufman, 2002).

Considerable data exists to suggest that measures of processing speed should not be used in place of the composite score to make a diagnosis. Broad cognitive speediness is operationally defined as “the speed with which simple stimuli are reproduced by an examinee” (Kamphaus, 2001). This factor is important to distinguish from other intelligence factors because of its relative emphasis on speed. The measurement of speed has always been controversial, from the earliest days of intelligence testing, when it was found to be a poor measure of psychometric “g” and poorly to moderately correlated with academic achievement (Kamphaus, 2001).

Table 1
*Predictive Validity of Part and Overall Composite Scores*

| | **Wechsler Individual Achievement Test-Second edition (WIAT-II)** | | | | |
|---|---|---|---|---|---|
| **WPPSI-III** | Reading | Math | Writing Language | Oral Language | Total Achievement |
| Verbal IQ | .60 | .56 | .59 | .72 | .77 |
| Performance IQ | .44 | .60 | .36 | .44 | .55 |
| PSQ | .31 | .55 | .41 | .39 | .36 |
| Full Scale IQ | .66 | .77 | .62 | .67 | .78 |
| GLC | .65 | .55 | .59 | .67 | .76 |
| **WISC-IV** | | | | | |
| VCI | .74 | .68 | .67 | .75 | .80 |
| PRI | .63 | .67 | .61 | .63 | .71 |
| WMI | .66 | .64 | .64 | .57 | .71 |
| PSI | .50 | .53 | .55 | .49 | .58 |
| FSIQ | .78 | .78 | .76 | .75 | .87 |
| **Stanford Binet Fifth ed. (SB5)** | | | | | |
| Fluid Reasoning | .59 | .57 | .47 | .58 | .65 |
| Knowledge | .60 | .72 | .51 | .63 | .71 |
| Quantitative Reasoning | .50 | .69 | .39 | .65 | .66 |
| Visual Spatial Reasoning | .53 | .69 | .41 | .71 | .69 |
| Working Memory | .50 | .62 | .33 | .62 | .61 |
| Nonverbal IQ | .52 | .72 | .42 | .70 | .70 |
| Verbal IQ | .75 | .79 | .58 | .78 | .83 |
| Full Scale IQ | .67 | .79 | .53 | .77 | .80 |

| | **Wechsler Individual Achievement Test – First edition (WIAT)** | | | | |
|---|---|---|---|---|---|
| **Reynolds Intelligence Assessment Scales (RIAS)** | Reading | Math | Writing | Language | Total Achievement |
| Verbal Intelligence Index | .67 | .67 | .61 | .70 | .73 |
| Nonverbal Intelligence Index | .43 | .46 | .43 | .35 | .41 |
| Composite Intelligence Index | .65 | .67 | .60 | .64 | .69 |
| Composite Memory Index | .55 | .59 | .55 | .50 | .58 |

| | **Basic Achievement Skills Individual Screener (BASIS)** | | | | | |
|---|---|---|---|---|---|---|
| | Age 7 | | | Age 11 | | |
| **Differential Ability Scales (DAS)** | Mathematics | Spelling | Reading | Mathematics | Spelling | Reading |
| Verbal | .49 | .41 | .56 | .34 | .46 | .60 |
| Nonverbal Reasoning | .54 | .52 | .62 | .66 | .36 | .49 |
| Spatial | .38 | .35 | .45 | .38 | .27 | .33 |
| GCA | .58 | .52 | .66 | .57 | .46 | .59 |
| Special Nonverbal Composite | .53 | .50 | .61 | .59 | .36 | .46 |

The relationship of processing speed to “g” has a long research tradition in individual differences inquiry (Deary & Stough, 1996) with one prominent finding -measures of reaction time and inspection time correlate moderately with measures of “g.” This, however, does not appear to be the case when considering the fourth factor of the WISC-IV-the Processing Speed Index (PSI). Despite the retention of its name from the WISC-III, the PSI (comprised of the Coding and Symbol Search core subtests) may involve less cognitively complex processing tasks in comparison to the typical processing speed models, thus weakening its relationship to "g." In fact, Carroll (1993) noted this distinction in task demands and assigned coding-like tasks to the stratum II factor of cognitive speediness, which is far removed from better measures of "g" at stratum II. Kranzler (1997) cited the modest loading of these component subtests as a rationale for not using the processing speed label adopted for this factor. In his own words he observed:

> Further research on what this factor measures is obviously needed, but in the meantime other names should be considered, because labeling this factor "Processing Speed" is inconsistent with the results of contemporary theory and research on the cognitive underpinnings of g and may mislead those unfamiliar with the literature (p. 114).

A recent study of this nature suggests the PSI factor measures motor skills more so than speed of cognitive processing, as the label suggests. Results of a validity study, based on test-criterion relationships, are provided in the WISC-IV Technical and Interpretive Manual (2003). Factor index screens for children aged 6-15 who were identified with significant motor delays or impairments were compared to a matched control group of children the same age. The mean score obtained on the PSI for the motor impaired group was well below their peers (PSI mean for Motor Impaired Group: 78.2; PSI mean for Control Group: 97.7), as well as below their scores on the three other indices (VCI: 95.5; PRI: 83.8; WMI: 92).

To further illustrate this point, consider the following study. Factor indices were used in an investigation (Glutting, McDermott, Prifitera, & McGrath, 1994) that employed a conjoint multivariate analysis of the WISC-III and the Wechsler Individual Achievement Test (Wechsler, 1992). Six prototypical profiles were obtained. The predominant distinction among these subtypes was overall level of intelligence. This study was designed to identify subtypes of WISC-III performance, exclusively on the basis of factor index scores, using cluster analysis. Two clusters were found that were differentiated primarily by different patterns of performance, with relative effectiveness on the PS factor being the most prominent (although not exclusive) source of variance. This result is also consistent with findings from previous studies. For example, in all six profiles that were found by Glutting et al. (1994) in the linking sample for the WISC-III and the WIAT, PS was consistently either the highest or the lowest of the four WISC-III factor index scores. The same was true in seven out of nine of the subtypes that were based on WISC-III subtest scaled scores in another investigation (Glutting, McDermott, et al., 1997). While variability on the PS factor index may provide important diagnostic information across the entire age range of the WISC-III, this score should not be used in place of the composite score to make a diagnosis.

In terms of its usefulness and reliability for predicting academic achievement, the WISC-IV PSI factor shows weaknesses similar to its predecessors. The PSI factor of the WISC-IV was designed to measure a child's ability to "quickly and correctly scan, sequence, or discriminate simple visual information" (Wechsler, 2003). For children aged 6:0-16:11, correlation coefficients between the PSI Composite Score on the WISC-IV and the Composite Scores on the WIAT-II (Reading, Mathematics, Written Language, and Oral Language) ranged from .46-.58. Correlations between the PSI factor of the WISC-IV and the Total Achievement Composite Score of the WIAT-II (for the same age group) ranged from .54-.60 (Wechsler, 2003). In contrast, the correlations between the WISC-IV's FSIQ, VCI, and PRI Scores and the WIAT-II Total Achievement Composite Scores yielded much stronger coefficients (.71-.87). These findings further support our original proposal which advises against using processing speed tests as a measure of general intelligence since it is a poor predictor and correlate of achievement in comparison to verbal and overall composite scores.

While a number of other well-known and widely used intelligence batteries include some type of processing speed measure, this subtest score is not an essential ingredient to derive an overall composite score (i.e. full scale IQ). For example, the Differential Ability Scale (DAS) includes a "core" battery, which is made up of four to six subtests, depending on the child's age. These subtests were selected because they were "the best measure of reasoning and conceptual abilities" that were available in the test battery (Kamphaus, 2001), and thus are considered to be most closely related to "g". A processing speed test is not included in this group of "core" subtests that are used to derive the General Conceptual Ability (GCA) score. Instead, the DAS offers a number of "diagnostic" subtests intended to measure "relatively independent abilities," of which the Speed of Information Processing test is included (Elliot, 1990). The Reynolds Intellectual Assessment Scales (RIAS) is yet another example of an increasingly-used measure of intelligence that does not include a specific measure of processing speed. In fact, to "Substantially reduce or eliminate dependence on motor coordination and visual-motor speed in the measurement of intelligence" is listed among the eight primary goals in the development of the RIAS (Reynolds & Kamphaus, 2003).

The Committee on Disability Determination for Mental Retardation (2002) of the National Research Council also suggested that composites with poor g-loadings should not be used as a substitute for a composite score in the diagnosis of mental retardation. They suggested a composite score should be preferred to any other scores. If the composite score is doubted as a valid measure of an individual's functioning, alternative index scores can be employed based compelling evidence of their g-loadings. They recommended using subtests that have high g-loadings such as crystallized ability (*Gc*), and fluid reasoning (*Gf*).

### *Practice Parameter 3*

*Visual/spatial/nonverbal, working/short-term memory, processing speed, and like scores may be used for screening purposes but not for making diagnostic decisions.* Our science-based approach is the most effective way of assessing general intelligence for the majority of people. While this technique works for most clients, there are cases where the use of the overall composite or verbal composite scores actually inhibits proper assessment. For example, what happens when the psychologist must determine the intelligence of a client who is bilingual, deaf, or has cerebral palsy? Is it fair to use the overall composite score or verbal composite score when measuring this person's intelligence if the person can't hear the questions, respond clearly, or speak English? How do we most accurately assess the general intelligence of these people? Is it even possible?

Consider the scenario where Kelly wants to buy a couch for her home. She measures exactly the space in her living room where the couch would go using a tape measure and finds she needs a couch 8 feet long or less. Kelly drives to the furniture store and sees the couch of her dreams. Naturally, she wants to make sure the couch will fit in her home before ordering. All she has to do is measure it. Oh no! Kelly realizes she left her tape measure at home, and of course no one at the store has one either. Kelly is distraught because she does not have a precise means of measuring the couch and is unable to predict with accuracy if it will fit in her living room.

Kelly tries to solve her problem three ways. First, she uses her human intuition and "eyeballs" the couch, making a guess as to its length. Kelly thinks the couch is about 13 feet. Her salesman thinks it's about 6 feet. Even using their best subjective judgments, Kelly and the salesman are not even in the same ballpark. Kelly's second option is to leave the store, go home and return with a tape measure. By doing this, however, Kelly risks losing the couch of her dreams to someone else. Kelly's third option is to use a less exact, though still somewhat reliable, means of measuring the couch and hope she is relatively close to the actual length. Under these less than optimal circumstances, Kelly decides this is her best option.

Kelly uses her own foot to measure the length of the couch. Kelly finds that the couch is 9 "Kelly feet" in length. She knows her foot is about 2 inches less than an actual foot so she subtracts 18 inches from her original measure to find that the couch is about 7.5 feet. Yea! It will probably fit in her living room.

True, Kelly arrived at a less precise result than she would have had she used a more accurate tool like a tape measure or yardstick. Still, using her foot was definitely more reliable than her or her salesman's intuition and certainly better than not measuring the couch at all.

Psychologists face a similar dilemma when trying to assess the intelligence of clients who enter the testing situation under special circumstances. For example, a psychologist is charged with assessing the intelligence of Fatuma, a child from Africa who is new to a school district. Fatuma has spoken Swahili for 10 years and English for 6 months. She likely will score vastly lower on the verbal composite than the performance composite score, thus skewing the full scale composite score and rendering it inaccurate for the purposes of assessing her general intelligence. Like Kelly, the psychologist must measure something (in this case intelligence) without the luxury of the best tool available (the overall composite score or verbal composite score). What can the psychologist do?

The psychologist could *guess* at Fatuma's intelligence, although as we saw in Kelly's case, human intuition is poorly correlated with actual outcomes (Dawes, 1995). Still, the psychologist needs to come up with a measure of Fatuma's intelligence now; she simply can not wait 5 years until Fatuma becomes academically proficient in English before assessing her (Cummins, 1981). The psychologist must find an alternate way to measure Fatuma's intelligence that is not adversely influenced by her inability to speak English. In this case, instead of the overall or verbal composite score, the psychologist could use a visual spatial composite score or another non-verbal composite score.

Nonverbal measures have become popular because of their potential benefits for assessing people with speech or language concerns such as new immigrants, bilinguals, and individuals with hearing or speech problems. In such cases, composite score and/or verbal/crystallized score may not accurately reflect their intellectual functioning. Instead, using nonverbal intelligence tests for this subgroup may be more reasonable. Although many traditional nonverbal measures may be measuring the Performance scale or Carroll's Stratum II Visualization/Spatial factor which has lower g-loadings than crystallized and fluid reasoning they still correlate somewhat with "g" and are much more accurate than subjectively guessing at one's intelligence, or not attempting to measure it at all (Kamphaus, 2001).

As an example, the General Abilities Measure for Adults ([GAMA], Naglieri & Bardos, 1997) is a nonverbal test that measures visual-spatial reasoning skills. Although GAMA has good potential as a nonverbal test, its reliability and stability are fairly weak and the IQs generated from the GAMA do not concur with the IQs generated from the comprehensive tests. It yields little relationship to verbal abilities as measured by other tests of cognitive ability (Kaufman & Lichtenberger, 2002). The possible use of visual/spatial/nonverbal score as a substitute for composite score remains controversial. Committee on Disability Determination for Mental Retardation (2002) suggested that visual/spatial measures can be used in place of a composite score. However, until further sound validation is available, visual/spatial/nonverbal scores are not considered to be good alternatives of composite scores.

Working/short term memory and processing speed index scores might provide useful information in inferring intellectual functioning for specific populations as well. For example, the Working memory and Processing speed indices in the Wechsler scales may yield lower scores than the Perceptual organization and Perceptual Reasoning index scores among adolescents and adults with learning disabilities. However, these patterns are not powerful enough to make a differential diagnosis and cannot be used to make a diagnosis of a learning disability (Kaufman & Lichtenberger, 2002). Rather the information from these scores can trigger further diagnostic or comprehensive tests to support or reject the initial observation. Donders (1995) demonstrated that processing speed scores may provide unique and valuable information about individuals with traumatic brain injuries. The Processing speed factor was found to correlate much more strongly with length of coma than did any other WISC-III factor or composite score, as well as any of the K-BIT standard scores.

Ideally, we believe that psychologists should use the overall composite score or verbal composite score when assessing general intelligence. Sometimes, however, circumstances and common sense forbid using these scores and instead dictate the use of composite scores less correlated with "g." Regardless of which scores are used when determining general intelligence, psychologists always should integrate all the available information to create the most accurate assessment possible (Kamphaus, 2001).

### *Parallels with the Past*

Ultimately, we have come "back to the future" of intelligence testing (Kamphaus & Kroncke, 2003) in that emphasis in virtually all intelligence test interpretation should be placed on the total score. The historical parallels between the research findings cited in this paper and the design and desires of the early intelligence testers are obvious. The forefathers of intelligence testing set a precedent by focusing on the total score, a model that has not been altered, only refined, over subsequent decades. Binet and Simon's stated purpose in 1905 of evaluating a level of "natural intelligence" separate from instruction remains a desired goal today. Their emphasis on assessing overall intellectual faculties and complex mental processes above and beyond simple component factors of intelligence provided a model to be followed by intelligence test developers throughout the last century.

As Kamphaus (1998) suggested, intelligence testing practice has not changed substantially since Binet's time. Many of the same tests and types of tasks included in Binet's initial scale remain in contemporary versions of intelligence scales today, modernized merely to reflect changes in lifestyle and linguistics. Though additional scores thought to represent distinct factors of intelligence were instituted with the publication of the Wechsler-Bellevue in 1938, the total score remained central to test interpretation. While research efforts to validate these distinct aspects of intelligence have proven moderately efficacious, the strongest evidence still remains for the overall score. This appears to be consistent with, rather than contradictory to, Wechsler's intentions. In 1974 Wechsler extrapolated on his conceptualization of the Verbal and Performance scales not as distinct abilities but as two "languages" by which general intelligence may be expressed. Considering the various scales proposed by various intelligence tests as means of <u>expression </u>of intelligence rather than distinct<u> types</u> of intelligence makes intuitive sense if one perceives intelligence to be an underlying latent trait that may be difficult to ascertain using only one assessment approach (e.g., verbal performance).

As emphasized throughout this paper, intelligence is a broad overarching construct that can be better utilized and discerned in its entirety rather than in fragmented parts. Spearman's conception of "g" as the most important intellectual trait, an underlying mental energy central to all intelligent problem solving, does not preclude the existence of specific factors but clearly supercedes them in terms of importance. Despite society's current inclination to deconstruct thoughts and ideas to minutia in an attempt to thoroughly understand their constitution, it seems evident that the most meaningful focus is on the Gestalt. In terms of intelligence, the whole appears to be greater than the sum of its parts.

We have come full circle in our understanding and conceptualization of intelligence test interpretation. Like the parable of the visually impaired individuals whose perception of an elephant was based on their limited assessment of its varied body parts, we too for a time lost sight of the whole by focusing on the component parts of intelligence. Though the hypothesized specific factors of intelligence (e.g., working memory, spatial ability) have merit, individually they do not accurately reflect the complex picture of intelligence. As John Godfrey Saxe, the author of the abovementioned parable, concluded, “Though each was partly in the right, they were all in the wrong.” We, too, are in the wrong if we attempt to compartmentalize intelligence, fragmenting it into parts that, in and of themselves, are less valid and conclusive than the total score.

## References

Binet, A., & Simon, T. (1905) Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L’Année Psychologique, 11*, 191-244.

Bloom, B. S. (1964). *Stability and change in human characteristics.* New York: Wiley.

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition. *Psychological Assessment, 10*(3), 285-291.

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*(2), 438-453.

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

Committee on Disability Determination for Mental Retardation. (2002). In D. J. Reschly, Myers, T. G., & C. R. Hartel (Eds.), *Mental retardation: Determining eligibility for social security benefits.* Washington, DC: National Academy Press.

Cummins, J. (1981) Age on arrival and immigrant second language learning in Canada. A reassessment. *Applied Linguistics, 2*, 132-149.

D’Angiulli, A., & Siegel, L. S. (2003). Cognitive functioning as measured by WISC-R. *Journal of Learning Disabilities, 36*(1), 48-58.

Dawes, R. M. (1995). Standards of practice. In S. C. Hayes, V. M. Follette, R. M. Dawes, & K. E. Grady (Eds.), *Scientific standards of psychological practice: Issues and recommendations.* Reno, NV: Context Press.

Deary, I. J., & Stough, C. (1996). Intelligence and inspection: Achievements, prospects, and problems. *American Psychologist, 51*, 599-608.

Donders, J. (1995). Validity of the Kaufman Brief Intelligence Test (K-BIT) in children with traumatic brain injury. *Assessment, 2*, 219-224.

Elliot, C. (1990). *DAS Handbook.* San Antonio: The Psychological Corporation.

Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, S. E., & Shaywitz, B. A. (1992). The validity of discrepancy-based definitions of reading disabilities. *Journal of Learning Disabilities, 25*(9), 551-561, 573.

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.), *Beyond traditional intellectual assessment: Contemporary and emerging theories, tests, and issues*. New York: Guilford Press.

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., McDermott, P. A., Prifitera, A., & McGrath, E. A. (1994). Core profile types for the WISC-III and WIAT: Their development and application in identifying multivariate IQ-achievement discrepancies. *School Psychology Review, 23*, 619-639.

Glutting, J. J., McDermott, P. A., Watkins, M. M., Kush, J. C., & Konold, T. R. (1997). The base rate problem and its consequences for interpreting children’s ability profiles. *School Psychology Review, 26*, 176-188.

Gustaffson, J. E., & Olav-Undheim, J. (1992). Stability and change in broad and narrow factors of intelligence from ages 12 to 15 years. *Journal of Educational Psychology, 84*(2), 141-149.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H. Brookes.

Hopkins, K.D., & Bibelheimer, M. (1971). Five-year stability of intelligence quotients from language and nonlanguage group tests. *Child Development, 42*, 645-649.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Hunter J. E., & Schmidt F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy and Law, 2*, 447-472.

Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence* (2nd ed.). Boston: Allyn & Bacon.

Kamphaus, R. W., & Kroncke, A. P.(2003). Back to the future of the Stanford-Binet. In G. R. Goldstein & S. Beers (Eds.), *Handbook of Psychological Assessment*. New York: Wiley.

Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003). The Flynn Effect and U.S. Policies: The impact of rising IQ scores on American society via mental retardation diagnoses. *American Psychologist, 58*(10), 778-790.

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

Kaufman, A. S., & Kaufman, N. L. (1990). *Manual for Kaufman Brief Intelligence Test (K-BIT)*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Lichtenberger, E. O. (2002). *Assessing adolescent and adult intelligence* (2nd ed.). Boston: Allyn & Bacon.

Kranzler, J. H. (1997). What does the WISC-III measure? Comments on the relationship between intelligence, working memory capacity, and information processing speed and efficiency. *School Psychology Quarterly, 12*(2), 110-116.

Livingston, R. B., Jennings, E., Reynolds, C. R., & Gray, R. M. (2003). Multivariate analyses of the profile stability of intelligence tests: high for IQs, low to very low for subtest analyses. *Archives of Clinical Neuropsychology, 18*(5), 487-507.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General Intelligence', objectively determined and measured". *Journal of Personality and Social Psychology, 86*(1), 96-111.

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "intelligent testing" approach to the WISC-III. *School Psychology Quarterly, 12*(3), 197-234.

Matheson, D. W., Mueller, H. H., & Short, R. H. (1984). The validity of Bannatyne's acquired knowledge category as a separate construct. *Journal of Psychoeducational Assessment, 2*, 279-291.

McDaniel, M. A., Schmidt F. L., & Hunter J. E. (1988). Job experience correlates of job performance. *Journal of Applied Psychology, 73*, 327-330.

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*, 163-175.

Moffitt, T. E., Caspi, A., Harkness, A. R., & Silva, P. A. (1993). The natural history of change in intellectual performance: Who changes? How much? Is it meaningful? *Journal of Child Psychology and Psychiatry and Allied Disciplines, 34*(4), 455-506.

Naglieri, J. A., & Bardos, A. N. (1997). *The manual of the General Ability Measure for Adults*. Minneapolis, MN: National Computer Scoring Systems.

Naglieri, J. A., Kamphaus, R. W., & Kaufman, N. L. (1983). The Luria-Das simultaneous–successive model applied to the WISC-R. *Jounral of Psychoeducational Assessment, 1*, 25-34.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77-101.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Psychological Corporation (1999). *Manual for the Wechsler Abbreviated Scale of Intelligence – WASI.* San Antonio, TX: Author.

Raguet, M. L., Campbell, D. A., Berry, D. T. R., Schmitt, F. A., & Smith, G. T. (1996). Stability of intelligence and intellectual predictors in older persons. *Psychological Assessment, 8*(2), 154-160.

Reynolds, C. R., & Kaufman, A. S. (1990). Assessment of children's intelligence with the Wechsler Intelligence Scale for Children–Revised (WISC-R). In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational* *assessment of children: Intelligence and achievement.* New York: Guilford.

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

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

Sameroff, A. J., Seifer, R., Baldwin, A., & Baldwin, C. (1993). Stability of intelligence from preschool to adolescence: The influence of social and family risk factors. *Child Development, 64*, 80-97.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Goff, S. (1988). The joint relation of experience and ability with job performance: A test of three hypotheses. *Journal of Applied Psychology, 37*, 407-422.

Shaywitz, B. A., Fletcher, J. M., Holahan, J. M., & Shaywitz, S. E. (1992). Discrepancy compared to low achievement definitions of reading disability: Results from the Connecticut longitudinal study. *Journal of Learning Disabilities, 25*(10), 639-648.

Shaywitz, S. E., Fletcher, J. M., Holahan, J. M., Shneider, A. E., Marchione, A. E., Stuebig, K. K., Francis, D. J., Pugh, K. R., Shaywitz, B. A. (1999). Persistence of dyslexia: The Connecticut longitudinal study at adolescence. *Pediatrics, 104*(6), 1351-1359.

Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The predictive validity of IQ. *Merrill-Palmer Quarterly, 47*, 1-41.

Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite: strengths and weaknesses. *Psychological Assessment, 16*(2), 133- 138.

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

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

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

Wechsler, D. (2003). *WISC-IV Wechsler Intelligence Scale for Children, Fourth Edition. Technical and Interpretative Manual.* San Antonio: The Psychological Corporation.