Philadelphia College of Osteopathic Medicine
DigitalCommons@PCOM

PCOM Psychology Dissertations

Student Dissertations, Theses and Papers

2016

Use of a Neuropsychologically Based Performance Model to Account for Adult Variation in Cognitive Test-Retest Performance

Ryan M. Murphy
Philadelphia College of Osteopathic Medicine, ryanmu@pcom.edu

Recommended Citation

Murphy, Ryan M., "Use of a Neuropsychologically Based Performance Model to Account for Adult Variation in Cognitive Test-Retest Performance" (2016). *PCOM Psychology Dissertations.* Paper 394.

Philadelphia College of Osteopathic Medicine

Department of Psychology

USE OF A NEUROPSYCHOLOGICALLY BASED PERFORMANCE MODEL TO ACCOUNT FOR ADULT VARIATION IN COGNITIVE TEST-RETEST PERFORMANCE

By Ryan M. Murphy

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Psychology

January 2016

PHILADELPHIA COLLEGE OF OSTEOPATHIC MEDICINE
DEPARTMENT OF PSYCHOLOGY

Dissertation Approval

This is to certify that the thesis presented to us by ___Ryan Murphy___ on the ___24___ day of ______July________________, 20_14___, in partial fulfillment of the requirements for the degree of Doctor of Psychology, has been examined and is acceptable in both scholarship and literary quality.

Abstract

The current study was designed to examine the appropriateness and effectiveness of a neuropsychologically based performance consistency model in comparison to traditional psychometric conceptualizations of reliability when examining test-retest performance on the Wechsler Adult Intelligence Scale – Third Edition (WAIS–III). Regardless of whether the sample was grouped by total sample or in reference to subject performance to the mean, an overall progression effect rather than regression to the mean was noted. When grouping subjects in terms of their relation to the mean, a poor goodness of fit obtained via chi-square analysis was found between traditional psychometric reliability estimates and actual results obtained via the performance consistency model. These findings are discussed in terms of the clinical utility of viewing changes in test-retest performance as potentially meaningful indicators of changes in cognitive functioning. This calls into question the appropriateness of previous research, which has typically eschewed the practice of subtest level interpretation of cognitive processing.

Table of Contents

Table of Contents

List of Tables

.

Chapter 1

Introduction

The duties of school psychologists are many and varied. They consult and problem solve with teachers, conduct group sessions, provide individual counseling, assist with systems level change, and design and monitor interventions. Perhaps the most traditional duty of a school psychologist is that of assessment. Although school psychologists utilize many assessment modalities, cognitive assessment is undoubtedly at the core of professional practice. This practice has been shown to be a theoretically and statistically sound method for assessing student strengths and weaknesses (Sattler, 2001).

However, traditional cognitive assessment also has its detractors. Among the arguments made are that intelligence tests fail to address aspects of multiple intelligences/learning styles, are culturally and/or racially biased, and take valuable time away from developing and monitoring interventions (Hale & Fiorello, 2001). In recent years, response to intervention (RTI) has arisen as a prominent issue in education. This is a tiered model that indicates children should receive successively more intense levels of intervention when evidencing academic or, to a lesser extent, behavioral difficulties. Although this model would reduce the number of children classified as learning disabled, it provides no insight into individual cognitive variations, assuming a lack of assessment of cognitive strengths and weaknesses. Providing interventions to children without this knowledge of their strengths and weaknesses is poor professional practice (Hale & Fiorello, 2004).

In the face of new approaches to school psychology practice such as RTI, the field of traditional cognitive testing needs to make assessment more meaningful in terms of linking assessment directly to intervention. Just as there are multiple opinions on the utility of cognitive testing, there are also various opinions on how to interpret cognitive test results. Very broadly speaking, there are those who believe that only global or overall/full scale scores should be interpreted and those who believe that meaningful interpretation of cognitive testing results lies at the composite and/or subtest level of analysis. Proponents of global/full scale analysis believe that only the overall score on a cognitive test should be interpreted. In composite/subtest level analysis, practitioners look for meaning in the scatter or profile of individual subtest scaled scores and standard scores on cognitive composites that comprise a cognitive battery. Furthermore, these practitioners believe that variation in subtest performance could indicate neurologic, learning, and emotional difficulties (Watkins, 2003).

Statement of the problem.

There are perceived inadequacies of the traditional psychometric model in offering realistic, clinically meaningful information about variation in test performance based on repeated administration of subtest tasks.

Purpose of the study.

The purpose of this study was to use Wechsler Adult Intelligence Scale – Third Edition (WAIS–III) standardization test-retest data to determine the extent to which a neuropsychologically based performance model fit the WAIS–III subtest test-retest data and how the decision consistency model of reliability data analyses compares with traditional psychometric correlational analyses. A variation of a decision consistency model that incorporates neuropsychologically based knowledge was used to examine expected Wechsler Adult Intelligence Scale – Third Edition (WAIS–III) subtest score patterns of change in performance from time 1 to time 2 and to offer a means to test the utility of this neuropsychologically based performance model for clinical practice.

Research questions.

Question 1: To what extent does a neuropsychologically based performance consistency model fit WAIS–III subtest test-retest data, and how do the data generated by a performance consistency model compare to the data generated by the traditional psychometric model ? Extrapolating from the findings in the neuropsychological literature, the following patterns of performance variation from time 1 testing to time 2 testing were hypothesized:

1. For a majority of cases, subtests that primarily involve retrieval from long-term storage or association with stored knowledge (Vocabulary, Information, Word Reasoning, and Similarities subtests) would yield score differences that reflect no change in performance or minor fluctuations in performance of -1 or +1 resulting from minor variations in cognitive efficiency; cases with changes would be biased toward a progression effect, that is, increases would outnumber decreases, even in situations in which regression to the mean would predict performance decreases or no change.
2. For a majority of cases, subtests that primarily involve the initial registration of and manipulation of verbal information in the mind (Digit Span Forward and Backward, Digit Span, Letter–Number Sequencing, and Arithmetic subtests) would vary in degree and frequency, depending on the specific nature of the subtest task. Subtests that involve the holding and manipulation of nonmeaningful, decontextual information (Digit Span Forward, Digit Span Backward, Digit Span) would distribute

relatively equally around a central tendency of no change, with relatively fewer but equal numbers of cases having positive and negative change both above and below the mean. Tasks that primarily involve the initial registration and manipulation of verbal manipulation that is more contextual and meaningful (Arithmetic and to a lesser degree Letter-Number Sequencing) would have a pattern of performance closer to that of tasks involving retrieval from long-term storage.

3. For a majority of cases, subtests that primarily involve novel problem solving (Matrix Reasoning, Picture Concepts, Block Design, and Picture Concepts subtest) or processing speed applied to simple but relatively novel tasks (Coding, Symbol Search, and Cancellation subtests) would yield score differences that reflect positive changes in performance, reflecting a greater progression effect than a regression effect. Score decreases would be similar in magnitude both in cases in which time 1 scores were above the mean and in cases in which time 1 scores were below the mean, due to greater effects of cognitive inefficiencies, thereby negating the effect of regression to the mean thought to be caused by random distribution of measurement error both above and below the mean.
4. When considered in total, the test-retest results would support an alternate model of performance change consistent with the neuropsychological literature on practice effects and cognitive efficiency and inefficiency, rather than a model of no change or fluctuations in the form of regression to the mean based on the traditional psychometric conceptualization of reliability.

Question 2: Does a neuropsychologically oriented performance consistency method offer any potential advantages over traditional psychometric methods in the type of information it provides test consumers regarding WAIS–III subtest performance patterns?

Chapter 2

Literature Review

This literature review incorporates a wide range of topics necessary for understanding the conceptual bases for this study. The review includes a discussion of clinical approaches to intelligence test score interpretation, comparing and contrasting global, composite score interpretation approaches with factor score and subtest score approaches; theoretical perspectives on the effects of aging on cognition; the specific cognitive capacities measured by the WAIS–III; the neuropsychology of practice-related changes in intelligence test performance; the traditional psychometric conceptualization of the construct of test reliability; the limitations of the traditional correlational procedures for determining test reliability; and alternative models for determining test reliability.

Global vs. Factor or Subtest Level Interpretation of Intelligence Test Performance

Full Scale IQ interpretation: Internal consistency and academic achievement. As noted in Chapter 1, one school of thought regarding interpretation of intelligence test scores is that only global or full scale IQ (FSIQ) should be interpreted when examining the results of cognitive testing. According to Watkins (2003), although subtest level analysis is a practice commonly employed by many psychologists, this interpretation is not appropriate for professional practice and lacks both clinical and statistical significance. Watkins explains further that while there is a multitude of evidence supporting interpretation of cognitive assessment results at the global level, psychologists have begun to apply interpretation to the subtest level despite the lack of evidence.

A review of the literature does indeed find support for the exclusive use of full scale IQ in interpretation. One argument for the appropriateness of full scale IQ interpretation is from a psychometric perspective. Watkins (2003) indicates that it is often incorrectly assumed that subtests are as psychometrically precise as full scale scores. Global IQ composites have stronger internal consistency reliability coefficients in comparison to subtest coefficients. For example, subtest reliability coefficients for the Wechsler Intelligence Scale for Children – Fourth Edition (WISC–IV; Wechsler, 2003) subtest reliability coefficients range from .79 to .90, with a median of .86 and index coefficients ranging from .88 for the Processing Speed Index to .97 for the full scale composite (Wechsler, 2003).

Another argument for the exclusive interpretive use of global intelligence quotients relates to academic achievement. Watkins and Glutting (2000) used regression analysis to examine the extent to which the Wechsler Intelligence Scale for Children – Third Edition (WISC–III, Wechsler, 1991) subtest variability predicted academic achievement in the areas of reading and math for both exceptional and nonexceptional students. The results of their study indicated that, for both learning disabled and non-learning-disabled students, subtest variability provided no meaningful prediction of reading and math scores beyond that predicted by the full scale scores. It was specifically noted that overall ability level accounted for more than half of the variance in performance on academic measures and that subtest profiles accounted for only an additional 5% to 8%. This led the researchers to conclude that to interpret full scale IQ is to act in congruence with scientific evidence, whereas to interpret subtest level performance is to act against scientific evidence. Further support for the power of FSIQ in predicting academic achievement is provided in a study conducted by Kahana, Youngstrom, and Glutting (2002). These researchers used student performance on the Differential Ability Scales (DAS) to show that although approximately 80% of students had a statistically significant intracognitive discrepancy, the overall ability score proved to be the most statistically useful predictor of academic achievement. They additionally found that intracognitive scatter added no further predictive power in determining academic achievement.

McGrew and Knopik (1996) examined subtest scatter in predicting academic achievement in the areas of reading, writing, and math. They specifically noted that although intracognitive strengths and weaknesses were common in the sample, these strengths and weaknesses were not found to be predictive of low academic achievement.

One study of note extended the findings of those mentioned above to a sample of students diagnosed with attention deficit/hyperactivity disorder (ADHD). Mayes and Calhoun (2007) examined the relationship between both WISC–III and WISC–IV performance and academic achievement. Congruent with previous research, the results indicated that, for both the WISC–III and WISC–IV, the score that accounted most for the variance in the prediction of academic achievement was the full scale IQ.

One argument that proponents of subtest level interpretation often use is that full scale IQ decreases in its usefulness and ability to predict achievement as the variability in factor and/or subtest scores increases (Flanagan & Kaufman, 2004). However, proponents of full scale IQ interpretation argue the opposite. Watkins, Glutting, and Lei

(2007) found that full scale IQ on both the WISC–III and WISC–IV was a statistically significant predictor of academic achievement in the areas of reading and math, whereas variability in factor scores was not found to make a statistically significant contribution to predicting achievement scores. In addition, this was true for both learning disabled and non-learning-disabled samples. Similarly, Daniel (2007) used simulation methodology to show that high levels of factor score variability are not uncommon among groups of individuals and that full scale IQ is an equally valid approximation of general cognitive ability for groups with high factor score variability as for groups with relatively uniform performance across factors.

Proponents of factor or subtest level interpretation of cognitive assessment results state that fluctuations in performance on factors or subtests is indicative of exceptionality (Watkins, 2003). However, a review of the literature in this area does not support this. Watkins (1999) utilized the WISC–III standardization sample to examine whether subtest variability was useful in distinguishing children who were classified as LD from the rest of the normative sample. The results of the study indicated that variability in subtest performance was not useful for diagnostic purposes in distinguishing students classified as learning disabled (LD) from those who were not LD. A similar result was found by Watkins and Worrell (2000), who used the WISC–III standardization sample to examine whether subtest variability, defined by the number of subtests that deviate by ±3 points from the mean IQ composite scores, had diagnostic utility in differentiating learning disabled from non-learning disabled students. They concluded that the observed presence of subtest variability lacked diagnostic utility for distinguishing students with learning disabilities.

A review of the literature also identified studies suggesting that profile interpretation is not useful in predicting emotional disturbance. For example, it was hypothesized by many that poor performance on certain combinations of subtests or large differences in performances between certain subtests is indicative of a variety of emotional difficulties (Watkins, 2003). One specific study, conducted by Beebe, Pfiffner, and McBurnett (2000), examined the contention that poor performance on the Comprehension and Picture Arrangement subtests of the WISC–III are related to deficits in social functioning. Results suggested that Picture Arrangement subtest performance was not indicative of difficulty in social functioning, and that performance on the comprehension subtest had only limited and inconsistent clinical significance in terms of its relationship to social functioning.

It has been suggested that poor performance on the Coding subtest of the WISC–III, caused by a decrease in the number of students correctly completing items as the subtest nears its time limit, may indicate difficulties in areas of social/emotional functioning, such as attention, distractibility, and motivation (Watkins, 2003). Although this may seem to be a plausible explanation for declining performance, Dumont, Farr, Willis, and Whelley (1998) found that interpretation of Coding subtest performance in isolation is inappropriate. Specifically, they measured the extent to which coding performance decreases as the time limit of the subtest approaches. They found that a decrease in correct item completion was found in most children and that poor coding subtest performance was not indicative of any disability, nor was it able to differentiate among various disabilities.

Some proponents of subtest level analysis have indicated that there are specific profiles of performance on subtests that are indicative of disabilities. For example, Kaufman (1994) indicated that children classified with disabilities, including ADHD and LD, frequently exhibit poor performance on the WISC–III Symbol Search, Coding, Arithmetic, and Digit Span subtests. This profile of low performance has become known as the SCAD profile. These subtests were thought to measure cognitive abilities, such as memory, auditory processing, sequencing, and visual-motor integration, as well as behavioral constructs, such as motivation and distractibility (Watkins, Kush, & Glutting, 1997a). A review of the literature of the utility of this profile yielded mixed results. Prifitera and Dersh (1993) found higher numbers of students with this profile in a sample of both learning disabled students and students with ADHD than was in the WISC–III standardization sample. To further explore this, Watkins, Kush, and Glutting (1997a) conducted a study that examined both the prevalence and diagnostic utility of this profile for a group of children diagnosed with both learning and behavioral/emotional needs. The authors found that although children diagnosed with disabilities had elevated SCAD profiles than children without disabilities, the profile was not useful in differentiating disabled from nondisabled students. Using the profile for such purposes resulted in inaccurate classifications. An additional finding of the study was that the SCAD profile was not significantly predicative of academic achievement.

Another subtest profile that has garnered attention in the interpretation of cognitive testing results is the ACID profile, which consists of the Arithmetic, Coding, Information, and Digit Span subtests of the WISC–III. As with the SCAD profile, the utility of this profile for diagnostic purposes has been contested (Frederickson, 1999).

Watkins (2003) noted inconsistent results of studies, with prevalence of the ACID profile in students with disabilities ranging from a 1% to 12%. Prifitera and Dersh (1993) found that, as with the SCAD profile, there was a higher rate of the ACID profile in children diagnosed with LD and ADHD than in the WISC–III standardization sample. Conversely, Watkins, Kush, and Glutting (1997b) found that, although the ACID profile is generally more prevalent in children with learning disabilities, it was not useful in differentiating disabled from nondisabled students. This finding is essentially the same for the SCAD profile study.

Thus, it appears that while various subtest profiles are observed more frequently in children with disabilities, the profiles do not have utility in differential diagnosis. It is also interesting to note that all of these studies were conducted using the WISC–III; further research is needed to determine if the findings extend to the normative and theoretical updates provided by the WISC–IV.

Proponents of analysis at the subtest level have attempted to extend the analysis to include recommendations for interventions based on subtest analysis. Watkins (2003) indicates that these interventions based on test score patterns are often referred to as aptitude by treatment interactions and involve matching cognitive aptitudes to instructional methodology. A review of the literature in this area indicates that although the practice may seem clinically logical, there is little empirical evidence to support it (Gresham & Witt, 1997). The researchers cite numerous studies indicating that aptitude by treatment interactions derived from modality matching, cognitive processing styles, and neuropsychological models have consistently failed to achieve statistical significance.

Subtest and factor level analysis: Statistical and methodological considerations. As noted above, proponents of exclusive full scale IQ analysis have conducted studies showing that the full scale IQ is the most important variable when predicting academic achievement and that factor and subtest level analysis contribute little, if any, additional information. However, Hale, Fiorello, Kavanagh, Hoeppner, and Gaither (2001) have identified methodological flaws in this research. They specifically noted that the studies supporting the exclusive use of full scale IQ in predicting academic achievement have all utilized hierarchical regression analysis. In this analysis, the variable that is entered first into the regression equation will subsume its own variance and the shared variance of subsequently entered variables (Hale & Fiorello, 2004). The result is that the variable entered first becomes the variable that is most predictive of the outcome measure. Proponents of full scale IQ entered the full scale score in first, followed by the factor/index scores. Thus, when analyzing the results, they concluded that only full scale IQ should be interpreted because it explained the majority of the variation in academic achievement, while the factors accounted for very little of the variance (Hale & Fiorello, 2004). If the factor scores had been entered first then they, and not the full scale IQ, would appear to be most predictive of academic achievement performance. Thus, the statistical methodologies are flawed.

Hale and Fiorello (2001) advocate the use of regression commonality analysis. This differs from traditional regression analysis in that it makes it possible to determine the amount of unique variance of a variable and the amount of variance that is common to all variables. According to the authors, this is accomplished by examining both unique and common variances, among all variables, separately. Using regression commonality analysis, the authors found that the WISC–III full scale IQ consisted primarily of unique variance, not shared variance. The authors indicated that for exclusive full scale IQ interpretation to be valid, commonality analysis would have to reveal that full scale IQ consists mostly of shared variance, but clearly indicated the opposite (Hale & Fiorello, 2001). Thus, they posited that cognitive functioning is comprised of at least four factors or unique aspects of intelligence, not one overall score.

It should be noted, however, that these researchers are not necessarily opposed to full scale IQ interpretation. They note that for children with little variability in performance across the factors that comprise the overall score, the full scale IQ may be the most parsimonious measure of cognitive functioning (Hale & Fiorello, 2004). However, they cited numerous studies that lend support to the concept that full scale IQ should never be interpreted when significant factor or subtest scatter is present in the profile of a child. They also note that significant variability in a child's cognitive profile does not necessarily mean that a disability is present. Rather, variability necessitates the interpretation of more specific cognitive capacities when attempting to gain a valid and useful conceptualization of individual strengths and weaknesses.

Factor level interpretation of intelligence test performance.

As noted above, just as there are those within the field who advocate the exclusive use of full scale IQ in interpretation of cognitive assessment results, there are also those who believe that the most valid indicators of the cognitive strengths and weaknesses of a child are found at the factor level of analysis and that factor level differences are not apparent when only interpreting the full scale IQ score.

Cross battery approach to assessment. The cross battery approach to assessment, also known as the CHC cross battery approach, offers a perspective of cognitive assessment interpretation that emphasizes interpretation at the factor level. This approach combines John Carroll's hierarchical model of intelligence and Horn and Cattell's multifactor model (Hale & Fiorello, 2004). This approach emphasizes that intelligence is not composed of only a single factor, but rather is composed of multiple clusters of cognitive abilities. Interpretation is focused on the results of assessment measuring both broad and narrow cognitive processes, including short term memory, long-term retrieval, quantitative knowledge, and correct decision speed (Flanagan & Ortiz, 2001). It also allows for examiners to identify specific cognitive strengths and weaknesses. This approach to assessment requires taking various subtests from different cognitive batteries and combining them into factors so that the various broad and narrow cognitive abilities may be assessed. According to Hale and Fiorello (2004), this approach to assessment has utility, and school psychologists should incorporate this approach into their daily practice.

Concordance–discordance model. Another related approach to interpreting subtest and factor level cognitive assessment results is the concordance–discordance model (Hale & Fiorello, 2004). This proposed model eschews the use of the traditional discrepancy model that looks for a significant difference between measured full scale IQ and areas of academic achievement when identifying children with a specific learning disability. Instead, these clinicians advocate three specific steps, involving interpretation of cognitive factor/index level scores, in interpretation of test results when determining if a student has a specific learning disability. The first step involves the identification of a concordance between a deficient academic achievement area and the cognitive processes theoretically linked to that area. The second step involves finding discordance between the deficient academic achievement area and the cognitive processing area not associated with the achievement area. The last step is to find discordance between the cognitive processing strengths and weaknesses (Hale & Fiorello, 2004). What is apparent in this model is the nonuse of full scale IQ in interpretation and the emphasis on factor/index scores. These authors also contend that this model of interpretation would ideally lead to more effective intervention development for the child because of the identification of specific cognitive strengths and weaknesses via examination of the factor/index scores. This is something to which the interpretation from exclusively the full scale IQ level would not be sensitive.

It should also be noted that the authors of this model indicate that concordance–discordance interpretation can only be conducted when there are no statistically significant differences on the subtests that comprise each factor. In situations in which performance on one subtest is significantly different from performance on another subtest within the same factor or index, the authors posit that a different factor score would have to be created. They specifically note that this could be accomplished by combining scores from subtests that cluster together both clinically and theoretically. It should be noted, however, that they emphasize the importance of using subtests with good psychometric properties and that, although this may not be ideal, it is preferable to using an index score that may misrepresent abilities due to variation in subtest performance within the index (Hale & Fiorello, 2004). Thus, this model appears to place great emphasis on factor and subtest level analysis and little, if any, emphasis on analyzing the full scale IQ.

Subtest level interpretation of intelligence test performance.

Despite the criticisms in the professional literature that were previously cited, the value of subtest level interpretation has been acknowledged in a very broad manner by the profession of psychology. In the *Standards for Educational and Psychological Testing* (1999), a publication coauthored by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement and Evaluation, the following statement appears regarding the importance of subtest level interpretation:

> Because each test in a battery examines a different function, ability, skill, or combination thereof, the test taker's performance can be understood best when scores are not combined or aggregated, but rather when each score is interpreted within the context of all other scores and assessment data. For example, low scores on timed tests alert the examiner to slowed responding as a problem that may not be apparent if scores on different kinds of tests are combined (page 123).

This quote also offers insight into why the professional literature contains so many more research studies eschewing the practice of subtest level interpretation rather than supporting it. The *Standards for Educational and Psychological Testing* note that subtest level interpretation needs to occur in the context of all the information gathered during an evaluation. When this occurs, clinicians can make well-informed inferences and generate strong hypotheses about variations in cognition and their impact on an individual's functioning in various settings. The research studies cited here did not evaluate the usefulness of subtest level interpretation in the context of all data gathered, but rather limited their investigations to analysis of subtest score patterns independent of the other information gathered in evaluations. Noting that subtest score patterns are not as effective as global IQ scores in predicting achievement test scores fails to address the real issue, the extent to which subtest level information can help to explain why an individual examinee may or may not be achieving at the level predicted by a global IQ estimate.

Process approach to assessment. Originally proposed by Edith Kaplan (1998), one perspective that proponents of subtest and/or factor level analysis embrace is known as the process approach to assessment. This perspective has also been referred to as the Boston process approach to neuropsychological assessment. This approach to interpretation posits that global scores based on right or wrong answers are not nearly as important as knowledge of the strategies that a student may use to obtain both correct and incorrect responses (Kaplan, 1998). Thus, observation of how a student approaches tasks during assessment is crucial. Similarly, McCloskey and Maerlender (2005) posit that the process approach is based on at least five related cognitive assessment principles: (a) identifying the specific processes that interact to produce an outcome leads to increased cognitive explanations for performance, (b) differences in the input, processing, and output demands for cognitive tasks can affect how a student performs on that task, (c) direct and systematic observation of how students engage in tasks are important to understanding how they perform tasks, (d) the mistakes that a student makes are as important as what they did correctly, and (e) direct observation can result in the creation of hypotheses about cognitive functioning.

A relatively recent development in the field of intelligence test construction was the application of the process approach to the WISC–IV, thus creating the WISC–IV Integrated (McCloskey & Maerlender, 2005). This assessment is a compilation of all of the subtests, both core and supplemental, of the WISC–IV, with the addition of 12 subtests specifically designed for the application of the process approach to assessment (McCloskey & Maerlender, 2005). Many of these process approach subtests are variations of the original WISC–IV subtests, but with alternative formats to assess how students approach a given task. Examples are multiple choice versions of the Similarities, Comprehension, and Block Design subtests.

In terms of how the process approach fits into the interpretation of cognitive assessments, McCloskey and Maerlender (2005) indicate that the process approach posits that using only the full scale IQ or the four factor scores in interpreting assessment results would miss important differences in performance in more narrow/specific cognitive tasks that are measured by subtests. Taken a step further, they also posit that even subtest level analysis may be too broad in some cases and may miss individual variations on the specific items that comprise subtests. Thus, individual item level analysis is also proposed by the process approach. This emphasis on interpretation at both the subtest and individual items levels is not surprising, considering the importance of this approach for understanding how students perform on specific cognitive tasks that are measured by specific subtests and items.

A review of the literature indicates that the process approach to assessment also has obtained support from various clinicians within the field. Berninger (1992) advocated for the use of the developmental neuropsychological perspective when assessing children. This theoretical perspective is closely related to the process approach and stresses the importance of identification of the profile of specific cognitive strengths and weaknesses when planning treatments. In addition, it posits that techniques such as subtest profile analysis are among the most appropriate techniques for determining these individual differences (Berninger, 1992). Further support for subtest level analysis is found in Berninger's views on academic functioning in areas such as reading and writing. She indicates that the traditional cognitive approach focuses exclusively on one construct, such as full scale IQ, that accounts for a majority of the variance in performance in these areas, whereas the process approach provides valuable insight into the various cognitive systems that interact with one another in completing tasks.

Similarly, Erickson (1995) lauds the use of the processing approach to assessment for a variety of reasons. He indicates that the traditional psychometric approach is too focused on what the individual achieves, and thus is deficient in examining the specific strategies used by the examinee in obtaining responses to cognitive tasks. He also contends that the traditional psychometric approach often lacks the error analysis that is crucial to the process approach and allows for increased utility for diagnostic and treatment considerations. However, he also offers some cautions for clinicians who have chosen to utilize the process approach to assessment, among them that much of the information gleaned from this approach is based on clinical rather than scientific or empirical evidence (Erickson, 1995).

Yet another theoretical perspective of the interpretation of cognitive assessment results that is similar to the process approach is the nomothetic versus ideographic approach explicated by Hale and Fiorello (2004). Nomothetic interpretation is normative based and focuses on levels of performance. Conversely, ideographic interpretation is individually based and focuses on a single examinee's pattern of performance. It is these patterns of performance, and how a child individually approaches those tasks, that the process approach to assessment relies so heavily upon.

Cognitive hypothesis testing model of assessment. Another model of assessment that utilizes a subtest and/or index level of interpretation when examining the results of cognitive assessment is the cognitive hypothesis testing (CHT) model (Hale & Fiorello, 2004). This model utilizes a combination of a traditional problem solving model, the assessment and confirmation/refutation of cognitive strengths and weaknesses, and developing interventions based on those cognitive strengths and weaknesses. In terms of determining cognitive strengths and weaknesses, the authors suggest the use of demands analysis and cite this process as the key to identifying disabilities and to treatment development (Hale & Fiorello, 2004). The authors indicate that conducting demands analysis consists of using subtest level analysis of the input, processing, and output demands of a task. In interpreting these demands analyses, they caution that the examiner must always remember that demands analysis is different for every child and that subtests can measure different things for different children. In addition, the model suggests an attempt to discern patterns of performance on the various cognitive tasks so that low performance on one particular processing demand across subtests would be identified as a suspected area of weakness and provide valuable information for treatment purposes. This level of interpretation is very different from the exclusive use of full scale IQ for interpretive purposes as previously discussed. It is also interesting to note the views of Hale and Fiorello on using fixed rather than flexible testing batteries for assessment purposes. Although they believe that fixed batteries are acceptable, they indicate that they prefer being able to choose and combine various subtests from a variety of batteries to measure various cognitive abilities. This, too, is not something that proponents of full scale IQ analysis would support. Thus, it is evident that the cognitive hypothesis testing model heavily relies on subtest level analysis in both determining the cognitive strengths and weaknesses of children and developing treatments based on this obtained information.

Reliability and classical test theory.

One of the major criticisms of subtest level analysis is that subtests are much less reliable than index or global IQ scores and therefore lack the psychometric characteristics needed for valid interpretation. To discuss this criticism, an understanding of the traditional psychometric concept of reliability is necessary, as is an understanding of alternate theoretical concepts of reliability and how they impact judgments about the reliability of subtest level interpretation.

The concept of reliability stems from classical test theory and, broadly speaking, refers to the extent to which variations in test scores reflect true differences in cognitive capacity, as opposed to random fluctuations in performance (Furr & Bacharach, 2008). It also has been broadly defined as the degree of consistency of scores obtained by a given test subject on repeated administrations of a measure on different occasions (Anastasi & Urbina, 1997). Salvia and Ysseldyke (2010) define reliability as the relative absence of random error during measurement. According to classical test theory, there are three properties and assumptions of all tests that form the basis for the theoretical conceptualization of reliability: observed scores, true score, and measurement error (Furr & Bacharach, 2008).

Observed scores, true scores, and measurement error. Understanding how these assumptions/properties are interconnected in classical test theory is essential to correctly conceptualizing the concept of reliability. Observed scores refer to the observed measurement value obtained by an individual on a test. Essentially, these are the scores obtained by test subjects and, according to classical test theory, are estimates of the true ability of a test subject (Furr & Bacharach, 2008). In contrast to observed scores, a true score is a reflection of the actual or real ability of a test subject. It should be noted that it is impossible to determine a person's true score; any score that is obtained by a test subject is only an estimate of the true score (Salvia & Ysseldyke, 2010). As noted above, the concept of reliability refers to how much variation in observed scores reflects or matches variations in true scores. The discord or inconsistency between observed scores and error scores is hypothesized to originate from the aforementioned concept of measurement error. More specifically, these are often unknown factors, such as characteristics of the measure itself, test administration variables, and test subject variables, that create differences in observed scores (Furr & Bacharach, 2008). These differences can either raise or lower observed score performance relative to the true score (Salvia & Ysseldyke, 2010). Error due to such unknown factors is known as random error. Random error should be differentiated from systematic error/bias, which consistently impacts a score in only one direction. An example of random error is a test administration error in which the rate of item presentation to the test subject was either faster or slower than indicated by the standardized directions. Conversely, an example of systematic error/bias is administering a reading comprehension assessment in which the content of the reading passages is about baseball, and the test subject has extensive background knowledge about the sport.

The degree of reliability for any given measure is directly dependent on two factors: how much of the difference in performance is due to real difference in the construct being assessed and how much of the difference in performance is attributable to measurement error (Furr & Bacharach, 2008).

Correlation coefficients. Anastasi and Urbina (1997) explain that the reliability of a measure can be represented by a correlation coefficient because all reliability is a measure of the amount of agreement between two test scores from separate administrations. Thus, the correlation coefficient represents the degree of the relationship between two scores. Correlation coefficients range from -1.0 to 1.0, with a correlation coefficient of 0.0 indicating no relationship between the variables and correlation coefficients of -1.0 and 1.0 indicating perfect negative and positive relationships, respectively (Salvia & Ysseldyke, 2010). Generally speaking, desirable correlation coefficients are around .80 or .90 (Anastasi and Urbina, 1997). Various methods can be used to calculate correlation coefficients, such as the Pearson product-moment correlation coefficient. This computation takes into account two main factors: the test subject's performance position in a group and the amount of deviation in performance with respect to the mean (Anastasi & Urbina, 1997).

Calculating reliability coefficients. In their conceptualization of reliability, there are at least four different ways to calculate reliability, based on classical test theory. An important distinction among these four methods is whether they conceive of reliability in terms of score correlations or proportion of variance. The other distinction is whether reliability is viewed as the relationship of observed scores to measurement error or the relationship of observed scores to true scores (Furr & Bacharach, 2008).

The first method involves conceptualizing reliability as the ratio of observed score variance to true score variance and is expressed by the following formula (Furr & Bacharach, 2008):

$$r_{xx} = \frac{S^2_T}{S^2_O}$$

A second method conceptualizes reliability as a lack of error variance and is expressed by the following formula (Furr & Bacharach, 2008):

$$r_{xx} = \frac{S^2_T}{1 - S^2_O}$$

A third method involves squaring the correlation between true scores and observed scores and is expressed by the following formula (Furr & Bacharach, 2008):

$$r_{xx} = r^2_{OT}$$

The last method conceptualizes reliability as a lack of correlation between observed scores and error scores and is expressed by the following formula (Furr & Bacharach, 2008):

$$r_{xx} = 1 - r^2_{OE}$$

Types of reliability. Because it is such a multifaceted concept, it is not surprisomg that a review of the literature reveals several different types of reliability, including test-retest reliability, alternate form reliability, interrater reliability, internal consistency, split-half reliability, Kuder–Richardson reliability, and coefficient alpha (Anastasi & Urbina, 1997; Furr & Bacharach, 2008; Salvia & Ysseldyke, 2010). When attempting to estimate reliability, it should be noted that no single method provides complete accuracy (Furr & Bacharach, 2008).

Test-Retest reliability. Given the nature of the present study, it seems most appropriate to begin the discussion of the various types of reliability with the test-retest method. This method involves comparing a person's observed performance on a measure with observed performance of the same person on the exact same measure after a second administration. The reliability coefficient when using this method is the correlation of time 1 and time 2 observed scores, and error variance equates to random performance variations from one administration to the other. The larger the reliability coefficient, the less likely it is that subject performance is being impacted by random error (Anastasi & Urbina, 1997).

This conceptualization of reliability includes the assumptions that true scores are stable across administrations and that the same amount of error variance is present on each testing occasion (Furr & Bacharach, 2008). The length of time between test intervals is a key consideration with this method, with longer time between intervals being more likely to change/challenge the necessary assumption of stability of true scores across administrations. Conversely, a short interval between testing might lead to an adverse impact on interpretation, due to carryover effects from the initial testing.

Interrater reliability. Interrater reliability, also known as examiner reliability or scorer reliability, refers to the degree to which behavior ratings from multiple raters are free from error variance associated with either the rater or examiner. Percentage agreement is the most common measure of interrater reliability (Sattler, 2001).

Alternate forms reliability. In this method of reliability, two or more forms of the same test are utilized in order to provide an estimate of reliability. With this method, the same subject is tested with one form on one occasion and a second form on subsequent assessment (Anastasi & Urbina, 1997). The obtained reliability coefficient represents the correlation between the observed scores on the two different versions of the test. In addition, these forms must both measure the same construct to the same extent and be standardized on the same population (Salvia & Ysseldyke, 2010). Similarly, in order to provide an appropriate estimate of reliability, the forms must be considered parallel. That is, they must have the same amount of variance, and they must measure the same set of true scores (Furr & Bacharach, 2008).

Internal consistency methods: Split half reliability and coefficient alpha. Yet another method to estimate reliability is known as internal consistency and does not require multiple forms of a test (Salvia & Ysseldyke, 2010). When conceptualizing reliability using internal consistency methods, two factors impact reliability estimates: the degree of consistency between the parts of the test and the length of the test. It is important to note that when applying this method, longer tests are more likely to have greater reliability than shorter tests (Furr and Bacharach, 2008).

One application of this method is known as the split half method and is accomplished by splitting the items of one test into two parallel subtests of equal size, calculating a composite score for each of these subtests, and then correlating the two composite scores. The formula for calculating the split-half method is known as the Spearman-Brown formula and mathematically represented as follows:

$$r = \frac{2r_{\mathrm{hh}}}{1 + r_{\mathrm{hh}}}$$

Another type of internal consistency reliability is known as coefficient alpha and is calculated by determining the mean split-half correlations based on all possible subtests of an assessment (Salvia & Ysseldyke, 2010). This method is closely related to a coefficient alpha method developed by Kuder–Richardson, which is coefficient alpha for test items that can only be scored as correct or incorrect (Sattler, 2001).

Coefficient alpha can be represented in two different ways: raw coefficient alpha and standardized coefficient alpha. Standardized coefficient alpha represents a reliability estimate for a test in which all items have been standardized, whereas with raw coefficient alpha, also known as Cronbach's alpha, the items are not standardized (Furr & Bacharach, 2008).

Relationship between reliability and the standard error of measurement. Another measurement of test error that is closely intertwined with the concept of reliability is known as the standard error of measurement *(SEM)*, and is defined as the mean standard deviation of error distributed around the true score of a test subject. The *SEM* provides information regarding how confident a clinician can be when interpreting observed scores (Salvia & Ysseldyke, 2010). The standard error of measurement also has been defined as the average size of the error scores; a larger SEM equates to decreased reliability because of the increased average distance/variation between true and observed scores (Furr & Bacharach, 2008). Anastasi and Urbina (1997) noted that the SEM is particularly useful when interpreting individual scores and often offers greater clinical utility than reliability coefficients.

Confidence intervals. Another concept based on reliability is that of confidence intervals. As noted above, it is not possible to know a person's true score. However, clinicians are not completely powerless in this regard. Specifically, confidence intervals reflect an estimation that a true score on an assessment is found within a specific range of scores (Salvia & Ysseldyke, 2010). Confidence intervals include both a range of scores within which the true score likely falls and the level of confidence with which we can estimate that a true score falls in that range. Generally, larger levels of confidence, expressed as percentages, are associated with larger score ranges and vice versa. For example, an FSIQ of 136 on the WISC–IV corresponds to a confidence interval of 130 to140 at 95%. This means that there is a 95% chance that the true score for the child falls between 130 and 140.

Factors impacting reliability. A review of the literature reveals multiple factors that impact reliability, such as test subject sample size, variations within testing sessions, guessing, variability of scores, test-retest interval length, homogeneity of test items, examiner characteristics, examinee characteristics, test length, consistency among parts of a test, and heterogeneity of test subjects (Furr & Bacharach, 2008; Sattler, 2001). The following section will discuss these factors.

Sample size. Reliability of a measure increases as sample size increases. This is because larger sample sizes yield smaller sampling errors associated with the reliability coefficient (Sattler, 2001).

Variation in the testing session. This factor indicates that fewer variations are associated with greater measurement reliability (Sattler, 2001). Examples might include a fire drill in the middle of a test session, loud noises from another room that are distracting to the examinee, or poor rapport with the examiner.

Guessing. According to Sattler (2001), when examinees respond randomly to items, it introduces error into the score and lowers reliability. It should be noted that this hold true even when guessing results in correct responses.

Variability of scores. Restriction of the range of possible scores has an impact on reliability. A small range of possible scores will result in lower reliability estimates, due to restriction of how much score can vary. A broad range of possible scores will produce higher reliability estimates due to the expanded possibilities for score variation (Sattler, 2001).

Test-Retest interval. Shorter intervals between test administrations are more likely to yield no change in performance and thus higher reliability estimates (Sattler, 2001).

Homogeneity of test items. This factor implies that more similar test items are likely to create more reliable measures (Sattler, 2001).

Test length. The greater the number of test items, the greater the likelihood of increased reliability for a measure (Furr & Bacharach, 2008).

*Consistency among parts of a test***.** Tests with greater internal consistency will have greater reliability. For example, when using the split-half method, greater consistency among the two subtests equates with greater internal consistency and thus reliability (Furr & Bacharach, 2008).

Sample heterogeneity. The greater the differences of examinee true scores, the greater the reliability. Stated another way, the greater the variability of a sample or group, the higher the reliability coefficient will be (Furr & Bacharach, 2008).

Examinee characteristics. Sattler (2001) reports numerous subfactors of examinee characteristics that can impact test reliability: overall ability level, test taking skills, ability to comprehend instructions and task requirements, mastery of specific test content, health, fatigue, motivation, affect, problem solving techniques, level of practice with test items, fluctuations in attention and memory, anticipation of test items, willingness to guess, and response to directions.

Examiner characteristics. It has also been noted that examiner characteristics can impact reliability estimates, such as errors associated with technique and style, personal needs, personal likes and dislikes, values, understanding of examinee, attention to the testing environment, ethnicity, administration and scoring errors, theoretical positions, recording techniques, bias, and interpretations (Furr & Bacharach, 2008).

Alternatives to the traditional psychometric theoretical conceptualization of reliability.

A review of the literature indicates several alternatives to the classical psychometric conceptualization noted above, such as domain sampling theory. According to Furr and Bacharach (2008), this theory is based on the assumption that test items are representative of a sample from a large indefinite number of potential test items. Parallel tests would hypothetically be created by selecting a number of items from a domain of items on two different tests. For example, if a new test were created from five test items from a domain measuring crystallized knowledge from one test, and a second test by taking five items measuring crystallized knowledge from another test, the tests should be parallel. Further, parallel test scores should correlate with each other. According to the authors, reliability is conceptualized as the mean correlation sizes among pairs of tests, with a certain number of items selected from the indefinite number of potential items, or domain (Furr & Bacharach, 2008).

Whitaker (2010) discusses two additional alternative approaches to the traditional psychometric theoretical conceptualization of reliability: generalization theory or G-theory (Cronbach, Rajaratnam, & Gleser, 1963) and item response (Rasch, 1960).

The logic underlying domain sampling theory forms the conceptual basis for G-theory (Furr & Bacharach, 2008). According to Roebroeck et al. (1993), in classical test theory, measurement error is estimated as a single component, and the traditional conceptualization of reliability does not allow the clinician to partition the measurement into different sources of error. They also report that classical test theory does not allow for generalization of the measurement error to studies with different designs. Conversely, G-theory is a broader approach to reliability that utilizes analysis of variance (ANOVA) to recognize and estimate multiple sources of error (Shavelson, Webb, & Rowley, 1989, as cited in Roebroeck et al., 1993) occurring in test-retest situations. As noted above, the locus of cognitive processing changes when the brain is exposed to the same information at a later time. By not taking into account these neuropsychological considerations, G-theory ignores implications for changes in performance from time 1 to time 2 assessment.

As noted above, yet another alternative to traditional psychometric approaches to reliability is known as item response theory (Rasch, 1960), in which reliability is based on the estimation of an underlying hypothetical trait theorized to impact test scores via a probability-based response model (Suen & Lei, 2007).

Additional procedures for estimating test reliability.

Decision consistency models. To account for some of the various shortcomings of the traditional psychometric approach noted above, McCloskey (1990) proposed using decision consistency models to obtain reliability estimations for criterion-referenced assessments, which are tests that measure subject performance in comparison to a specific criterion. The use of data consistency models has been reported in the literature by multiple sources (Subkoviak, 1980; Traub & Rowley, 1980; Van Der Linden, 1980 as cited by Whitaker, 2010). An example of criterion referenced assessment is a test measuring multiplication skills in which raw scores of 50 or above would be considered passing, and raw scores below the criterion, i.e., 50, would be considered failing.

Traub and Rowley (1980) note one of the main differences between reliability estimates for criterion and norm-referenced assessments lies in the relative importance of raw score variability. In norm-referenced assessments, this variability is given exclusive importance, whereas in criterion-referenced assessment, more emphasis is placed on the ability of the test across multiple administrations to give similar classifications based on the criterion being measured (Whitaker, 2010). Using the previous example, if students passed both trial 1 and trial 2 of the multiplication test with regularity, the instrument could be said to be reliable.

Utilization of a decision-consistency model consists of calculating the percentage of agreement for test subjects who receive the same classification per the criterion or cut score (McCloskey, 1990; Whitaker, 2010).

Use of decision-consistency models with norm-referenced assessments. Whitaker (2010) noted several examples of using decision-consistency models to predict reliability with norm referenced assessments. In one such instance, a decision-consistency model was applied to estimate reliability for specific scores obtained from the Kaplan-Baycrest Neurocognitive Assessment (KBNA), a measure used to assess various areas of neuropsychological functioning, including immediate and delayed memory, reasoning and conceptual shifting, spatial processing, verbal fluency, and attention. Leach, Kaplan, Rewilak, Richards, and Proulx (2000) utilized a decision-consistency model to indicate the percentage of test subjects whose criterion (qualitative classification range) did not change from time 1 to time 2.

Another example reported by McCloskey (1990) compared interrater reliability estimates for two early childhood behavior rating scales using the traditional psychometric approach with the correlation method and an agreement grid conceptually based on decision-consistency models. This grid was created using three different agreement percentages. The first, known as an identical ratings percentage, depicted the percentage of exact agreement from time 1 to time 2 (rating 1 to 2). The second, known as an increased ratings percentage, depicted percent increase in ratings from the first to the second rating. The last was a decreased ratings percentage that depicted percent decrease from time 1 to time 2.

McCloskey (1990) then compared reliability results using both correlation and the agreement grid for the first rating scale. The correlation method yielded a coefficient of .59. As noted in the above interpretive guidelines, this would indicate only a moderate degree of reliability. Conversely, the percent of agreement calculations found a much higher level of agreement, specifically 81.6% overall for identical ratings between time 1 and time 2 (McCloskey 1990). Restricted range of ratings was cited as the primary reason for this difference because range restriction adversely impacts correlational calculations, but did not impact decision-consistency model agreement percentages.

On the second rating scale, both the identical ratings percentage (76%) and correlation coefficient (.84) were high. McCloskey (1990) attributed the high percentage of increased ratings (21.5%) to either the tendency of responders to report higher ratings over time without possible increases in behavior or expectancy effects. Overall, this study indicated various advantages of the agreement grid, and thus decision-consistency models, in comparison to traditional correlational procedures. One such advantage involves interpretive considerations. Specifically, the grid was able to provide information about the impact of expectancy effects, as well as the nature of test-retest score disagreements (Whitaker, 2010). Stated another way, the agreement grid is more apt to provide information on similarities or dissimilarities between time 1 and time 2 scores. Correlational procedures were not able to accomplish this. The other advantage noted by McCloskey (1990) is that regardless of test score distribution, the decision consistency model still was able to yield accurate information concerning the degree of agreement in test-retest scores.

Whitaker (2010) concluded that McCloskey's modified decision-consistency model could be applied to the interpretation of WISC–IV scores, specifically by examining the degree of consistency or inconsistency at the subtest level for time 1 versus time 2 performance. This approach would provide valuable interpretive information in comparison to traditional correlation approaches in that it does not obscure patterns of performance in score changes from time 1 to time 2. This additional information can be obtained by comparing hypothesized results based on classical test theory such as practice effects and true score estimates, with the modified decision-consistency model calculations for amounts of positive change, no change, and negative change.

Cognitive capacities measured by the WAIS–III subtests.

The WAIS–III (Wechsler, 1997) is designed for use with individuals ages 16 through 89. The instrument is comprised of 14 subtests, 11 core and 3 supplemental. Exploratory factor analysis of the WAIS–III yielded a four factor model: Verbal Comprehension, Perceptual Organization, Working Memory, and Processing Speed (Sattler, 2001). Verbal Comprehension represents verbal-related ability including both verbal knowledge and verbal reasoning. Perceptual Organization refers to performance-related ability and mental processes such as nonverbal reasoning, attentiveness to detail, and visuomotor integration. Working Memory represents a memory-related ability whereby a person holds information in mind so that operations or manipulations can be performed. Lastly, Processing Speed refers to perceptual processing and psychomotor speed (Sattler, 2001). In addition to these indexes, the WAIS–III subtests may be organized in the traditional manner into the Verbal Scale and Performance Scale.

The Verbal Comprehension Index consists of three core subtests, Vocabulary, Similarities, and Information, and one supplemental subtest, Comprehension. The Perceptual Reasoning Index also consists of three core subtests, Picture Completion, Block Design, and Matrix Reasoning, and one supplemental subtest, Picture Arrangement. The Working Memory Index includes two core subtests, Arithmetic and Digit Span, and one supplemental subtest, Letter-Number Sequencing. The Processing Speed Index is comprised of two core subtests, Coding and Symbol Search, and one supplemental subtest, Object Assembly. Table 1 provides a summary of the WAIS–III core and supplemental subtests.

Table 1

Description of WAIS–III Core and Supplemental Subtests

Subtest	Description
Picture Completion	A set of color pictures of common objects and settings, each of which is missing an important part that the examinee must identify
Vocabulary	A series of orally and visually presented words that the examinee orally defines
Digit Symbol-Coding	A series of numbers, each of which is paired with its own corresponding hieroglyphic-like symbol. Using a key, the examinee writes the symbol corresponding to its number
Similarities	A series of orally presented pairs of words for which the examinee explains the similarity of the common objects or concepts they represent
Block Design	A set of molded or printed two-dimensional geometric patterns that the examinee replicates using two-color cubes
Arithmetic	A series of arithmetic problems that the examinee solves mentally and responds to orally
Matrix Reasoning	A series of incomplete gridded patterns that the examinee completes by pointing to or saying the number of the correct response from five possible choices
Digit Span	A series of orally presented number sequences that the examinee repeats verbatim for digits forward and in reverse for digits backward
Information	A series of orally presented questions that access the examinee’s knowledge of common events, objects, places, and people

(continued)

Subtest	Description
Picture Arrangement	A set of pictures presented in a mixed-up order that the examinee rearranges into a logical story sequence
Comprehension	A series of orally presented questions that require the examinee to understand and articulate social rules and concepts or solutions to everyday problems
Symbol Search	A series of paired groups, each pair consisting of a target and a search group. The examinee indicates, by marking the appropriate box, whether either target symbol appears in the search group
Letter Number Sequencing	A series of orally presented sequences of letters and numbers that the examinee simultaneously tracks and orally repeats, with the numbers in ascending order and the letters in alphabetical order
Object Assembly	A set of puzzles of common objects, each presented in a standardized configuration, that the examinee assembles to form a meaningful whole

Several authors (McCloskey, 2009; Miller, 2007; Whitaker, 2010, Miller & Hale, 2008) have identified both primary and secondary cognitive and neuropsychological processes likely assessed by the Wechsler subtests. These authors note that subtests measure both the construct, the primary cognitive process, and the secondary cognitive processes that, while not necessarily the goal of assessment, nonetheless impact performance. Variations in subtest performance can originate from efficient use of these primary and secondary capacities. These capacities are: executive functions, memory functions, auditory perceptions, language functions, reasoning ability, visuomotor processing speed, visual perception, and visual processing speed.

Executive functions have been defined as cognitive processes that serve a directive role by cueing the use of other mental capacities and coordinating multitasking efforts (McCloskey, 2009). Similarly, Miller (2007) notes that executive functions command and control other cognitive processes. When conceptualizing executive functions, it is important to remember that poor subtest performance may not necessarily be a result of a deficit in a primary capacity, but rather reflect difficulties with executive functioning. For example, the Vocabulary subtest is primarily a measure of word knowledge. However, executive functions allow subjects to retrieve that knowledge from long-term storage. Viewed this way, poor performance on the subtest could be the result of poorly developed word knowledge, deficits in executive functioning, or some combination of the two.

McCloskey (2009) notes that memory functions encompass a broad spectrum of capacities such as initial registration of information, retrieval from long-term storage, holding and manipulating information in mind and extending the immediate moment into the future. Miller (2007) presents a five-factor model of memory functions: immediate memory, long-term (delayed) memory, associative memory and learning memory, working memory, and semantic memory. Within McCloskey’s (2009) framework, auditory and visual perception are basic cognitive processes, narrow-band cognitive capacities that organize sensory input, either auditory or visual. Whitaker (2010) notes that auditory input consists of elements such as the perception and discrimination of speech sounds, as well as comprehension of grammar and syntax.

Language processes refers to a group of cognitive abilities that consist of phonological processing, receptive language, and expressive language (Miller, 2007).

McCloskey (2009) notes that reasoning is an ability that refers to a broad-band cognitive capacity that constrains memory and learning and operates on mental representations created via information input at the basic processing level. Processing speed is the speed with which one or more basic processes can be coordinated and applied (McCloskey, 2009). Visual processing speed then refers to coordinating and applying these basic processes when given visual stimuli. Miller (2007) notes that speed and efficiency of cognitive processing is impacts all cognitive capacities.

Neuropsychological conceptualizations of practice-related changes in intelligence test performance.

Traditional approaches to test reliability posit that scores should be stable from time 1 to time 2 and that changes is scores from time 1 to time 2 reflect measurement error. Alternately, however, performance on repeated administrations of a task has been examined within the neuropsychological literature. This has yielded findings that are not consistent with the traditional position that change in performance from time 1 to time 2 results from undesirable measurement error (Whitaker, 2010). The origin of these changes or gains in performance is hypothesized to be a concept known as the practice effect, whereby performance gains on repeated measures is thought to be the result of increases in cognitive processing efficiency due to factors such as prior exposure to the task, learning, or intervention (Sattler, 2001; Whitaker, 2010). Although classical test theory acknowledges the occurrence of practice effects, these effects are viewed as undesirable sources of variation that are associated with measurement error rather than real differences in changes in brain function that follow predictable patterns.

Koziol and Budding (2010) note that the presence of improved performance via practice effects illustrates that test subjects have learned, remembered, and applied information to subsequent administrations. This conceptualization is consistent with what is known about neuropsychological changes that take place when brains are exposed to information. These authors also differentiate two types of practice effects: content practice effects and procedural practice effects. Content practice effects are gains in performance once being exposed to the content or makeup of material and are particularly dependent on the medial-temporal lobe memory system. Conversely, procedural practice effects are related to learning a specific cognitive skill required for successful completion of a task (Koziol & Budding, 2010). Whitaker (2010) noted that practice effects are likely to be more pronounced for novel tasks encountered a second time than for familiar tasks that measure crystallized knowledge.

Before exploring current theories about cognitive processing difficulties, it is useful to examine a historical perspective. Goldberg (2009) indicates that although hemispheric specialization has long been a tenet of neuropsychological interpretation, this is an oversimplistic and often incorrect conceptualization. For example, the traditional notion in linking language to the left cerebral hemisphere and spatial processing to the right cerebral hemisphere is only partly correct.

More recently, instead of trying to localize specific brain areas or hemispheres to specific actions, researchers have begun to conceptualize the location of cognitive processing by the relative novelty or familiarity of tasks. The genesis of this relatively new conceptualization was a theoretical paper published by Goldberg and Costa (1981), which posited that the left hemisphere is more adept at processing familiar information that is routine, whereas the right hemisphere is more adept at processing novel information. This is known as the novelty-routinization hypothesis.

Goldberg (2009) defined learning as both change and a transition from inefficient behavior to more effective behavior. Further, initial stages of learning are characterized by novelty, whereas toward the end of the learning process, stages become much more familiar or routine. Goldberg also hypothesizes that the structural and chemical differences observed between the right and left hemispheres are a product of evolution to ensure the learning process as information moves from novel to routine.

Because the right hemisphere is more adept to processing novel information, and the left hemisphere more adept at processing routinized information, it is not surprising that Goldberg (2009) argues a right-to-left information transfer. He notes that mental representations develop in both hemispheres, but their rates of formation differ, with the right hemisphere more responsible for early learning and the left hemisphere more responsible during later stages of learning. This is accomplished by a gradual shifting in cognitive control (Goldberg, 2009).

However, this is not to say that tasks are specific to a particular brain location, but that the right and left hemispheres are interconnected, and the relative degree of right versus left hemispheric involvement is heavily dependent on the nature of the psychological processes involved in completing the task (Hale & Fiorello, 2004).

The genesis of the novelty-routinization hypothesis by Goldberg and Costa (1981) is the earlier writing of Luria (1973). Luria’s work posits the presence of three functional units in the brain. The reticular activation system, along with its related structures that are implicated in motor tone maintenance and activities such as walking, comprises the first functional unit (Hale & Fiorello, 2004). The posterior occipital, parietal, and temporal regions of the brain comprise the basis for Luria's second functional unit and are devoted to receiving, storing, and analyzing information. Lastly, the frontal lobes comprise Luria's third functional unit and are heavily implicated in cognitive regulation (Hale & Fiorello, 2004). According to Majovski (1997) the internal representation and organization of information was the basis upon which Luria posed right-left cerebral specialization. Similarly, Hale and Fiorello (2004) note that hemispheric division of labor is based on the specific neuropsychological process required for the task, as opposed to the stimulus itself. For example, Goldberg (2009) reports studies in which, regardless of whether stimuli are verbal or visual, novel information was processed in the right hemisphere and routinized information in the left. This runs contrary to the traditional notion of the left –verbal, right–visual dichotomy.

In addition to right-left differences, Luria also noted a gradual shift from anterior to posterior regions of the brain when acquiring cognitive skills (Luria, 1973). Specifically, the learning of new and novel information is subsumed by the anterior regions of the brain, notably the frontal lobes. Conversely, posterior regions of the cerebral cortex are more prominently involved in performing previously learned and mastered skills (Goldberg, 2009).

Recent advances in technology have made possible exponential increases in the knowledge of how the brain works and processes information. Among these advances are two neuroimaging techniques known as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET). These techniques are based on the premise that neural activity is correlated with cerebral blood flow levels (Posner &

Raichle, 1994; Roland, 1993). As described below, numerous studies have utilized these methods to examine neuroanatomical changes to the location of cognitive processing when performing various tasks.

Neuroimaging evidence for a right-to-left hemispheric transition of cognitive processing. Berns, Cohen, and Mintun (1997) utilized PET to examine the brain's response to novelty. Participants in their study performed a simple reaction-time test for two different conditions. In general, the study found that as examinees became more familiar with a task, i.e., trained, cognitive processing shifted from right to left. The study included a learning/training condition, and following this training, increased cerebral blood flow was found in the left premotor, left anterior cingulate, and right ventral striatum; decreased blood flow was observed in the right dorsolateral prefrontal and parietal areas.

Newman-Norlund, van Schie, van Zuijlen, and Bekkering (2007) utilized fMRI to assess the role of the mirror neurons in complementary actions. In general, they found that engagement in complementary actions that were different, i.e., more novel, was associated with greater activation in the right inferior frontal and right inferior parietal than in imitative actions, which were more routine.

The learning of novel visuomotor tasks was assessed by Staines, Padilla, and Knight (2002) by using event-related potentials characterizing neural activity associated with sensorimotor processes. Results of the study noted a two-dimensional shift following practice of the task, with cognitive processing moving from the right to left hemispheres, but also from the frontal lobes to the parietal lobe. This right-left and anterior-posterior shift is consistent with the literature noted above.

Another two dimensional right-left, front-back shift was noted in a study conducted by Kamiya, Umeda, Ozawa, and Manabe (2002, as cited by Goldberg, 2009). Using electroencephalogram frequency, these authors found that as familiarity with a task increased, there was a right-to-left and anterior-to-posterior change in cognitive processing. During initial exposure to a task, right frontal regions were especially active. Halfway through exposure to the task, right frontal involvement shifted to left frontal involvement and was accompanied by both right and left posterior activation. During the end of the task, little frontal involvement was noted, but left posterior regions displayed greater activation (Goldberg, 2009).

Raichle et al. (1994, as cited by Whitaker, 2010) examined practice-induced performance improvements by using PET to investigate neural activation patterns during both naïve and practiced performance of a verbal-response selection task. The naïve condition consisted of participants being asked to name an acceptable verb when visually presented with a noun. Conversely, during the practice condition, a different group of subjects were given 15 minutes to rehearse self-generated noun–verb associations learned after being asked to identify a verb to match with nouns on a list.

The results of the study indicated differential activation patterns based on each condition. Anterior cingulate, left prefrontal, and left posterior temporal cortices and right cerebellar cortical activation characterized the naïve condition. Although this pattern was again partially found during the novel condition, it was almost completely absent during the practice condition. An additional finding of the study was a significant increase in cerebral activation in the left medial occipital region for the practice condition in comparison to both the naïve and novel conditions.

Additional neuroimaging studies addressing general changes in cognitive processing. Seger (2006) defines sequence learning as learning a sequence of events over time. Within the area of sequence learning, Koziol, Budding, and Chidekel (2010) reviewed multiple studies that implicated activation in regions of the frontal lobe, the head of the caudate nucleus, and the anterior putamen. These authors further noted that there are changes found over time in the learning of motor skills. Poldrack et al. (2005) compared both initial and subsequent performance on the sequential reaction time task (SRT) and noted that initial performance of the task was associated with striatal, parietal, and frontal activation and decreases in activation for the putamen, globus pallidus, and supplementary area of the frontal cortex following training and subsequent task performance.

Beauchamp, Dagher, Aston, and Doyon (2003) used a modified version of the Tower of London test (Shallice, 1982) to examine the planning performance of 12 subjects. Participants were scanned on four occasions using PET. During initial performance of the task, high levels of activation were noted in the dorsolateral prefrontal, orbitofrontal, left parietal cortices, caudate nucleus, cerebellum, and premotor cortex. Conversely, decreases in activation in the medial orbitofrontal and frontopolar cortices were noted during later administration, after learning had occurred (Beauchamp et al., 2003). Similarly, Hubert et al. (2007) utilized the similar Tower of Toronto test to examine phases of procedural learning. They noted initial activation in the prefrontal cortex, cerebellum, and parietal regions. During a second condition, activation was noted in the occipital lobes, the right thalamus, and the caudate nucleus. Finally, during a third condition, activation was noted in the left thalamus and anterior lobes of the cerebellum

(Hubert et al., 2007). In conclusion, Seger (2006, as cited by Koziol, Budding, & Chidekel, 2010) indicates a general pattern in which as procedural learning increases, there is a shift in the seat of cognitive processing from executive and visual corticostriatal loops to the motor loop. Poldrack and Willingham (2006) supported this, citing multiple studies indicateing the prominence of the role of the prefrontal cortex during initial stages of learning, shifting to supplementary motor areas for later skill acquisition.

Intelligence test-retest practice effects. Given what has been discussed above about hemispheric changes in cognitive processing as information becomes less novel, it is not surprising that cognitive assessment is particularly sensitive to the impact of practice effects on subsequent administrations. Also, as noted above, the shorter the test-retest interval, the greater the hypothesized practice effects. Meta-analytic study revealed that research has repeatedly shown that subjects make gains in test performance across repeated administrations (Hausknecht, Day, & Thomas, 2006). Tuma and Appelbaum (1980) utilized the Wechsler Intelligence Scale – Revised (Wechsler, 1974) to examine the impact of practice effects over a 6-month period in 45 children ranging in age from 7.8 to 15.0 years. Although Verbal IQ stability estimates did not reveal much change, Performance IQ and FSIQ displayed significant practice effects, with mean gains of 8 and 5 points, respectively (Tuma & Appelbaum, 1980). This study supports the above assertion that gains in practice effects are likely to be seen on more novel information, in this case the performance subtests.

Ryan, Glass, and Bartels (2010) utilized the Wechsler Intelligence Scale for Children – Fourth Edition (WISC–IV; Wechsler, 2003) to examine test-retest stability in 43 elementary students tested approximately 11 months apart. Results from their study indicated that stability coefficients ranged from .26 for Picture Concepts to .84 for Vocabulary. At the composite level coefficients ranged from .54 for Processing Speed to .88 for Full Scale IQ. They further noted that only the Vocabulary subtest and Full Scale IQ had levels of stability; for all remaining areas, the standardization sample yielded significantly larger coefficients than did the 43 subjects from the study. Finally, the authors concluded that although the average practice effects were not significant, the range of gain or loss for some subjects was large (Ryan, Glass, & Bartels, 2010).

Watkins and Smith (2013) also utilized the WISC–IV to examine stability of test performance. However, this study was conducted over a longer period (mean interval of 2.84 years) and with a larger sample (344 students). The results indicated index level reliability estimates ranging from .65 for Processing Speed to .76 for Perceptual Reasoning. Full Scale IQ reliability estimate was .82. In general, subtest level reliability coefficients were lower than composite level measures. The authors further noted that 29%, 39%, 37%, and 44% of students had Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed scores, respectively, that varied by 10 or more points; 25% of students had FSIQ scores that varied by more than 10 points. The authors therefore concluded that clinicians should not assume long-term stability in scores across long test-retest intervals (Watkins & Smith, 2013).

Kaufman (2003), Lezak, Howieson, and Loring, (2004), and Shatz, (1981) have examined the possible impact of variables such as age, time elapsed between assessment intervals, motor speed requirements, and the novelty of tasks. Other researchers administered the WAIS (Wechsler, 1955) to college students at intervals ranging from 1 week to 4 months to examine the concept of practice effects. Results indicated that the greatest gains in performance at both the composite and full scale IQ level were noted at the 1-week interval. Conversely, the smallest gain in performance was noted at the 4-month level (Catron and Thompson, 1979).

The differential impact of practice effects on the Wechsler scales was examined by Kaufman (2003). Test-retest data for the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967), Wechsler Preschool and Primary Scale of Intelligence – Revised (WPPSI–R; Wechsler, 1989), Wechsler Intelligence Scale for Children – Revised (WISC–R; Wechsler, 1974), Wechsler Intelligence Scale for Children – Third Edition (WISC–III; Wechsler, 1991), and Wechsler Adult Intelligence Scale – Revised (WAIS–R; Wechsler, 1981) revealed the highest median standard score gains for the Performance IQ (median gain of 9.0 standard score points), followed by the FSIQ (6.8 standard score points), and Verbal IQ (3.2 standard score points) (Kaufman, 2003).

Whitaker (2010) noted a similar pattern of greatest practice effect gains on novel tasks as opposed to crystallized knowledge tasks on other tests of cognitive abilities, such as the Kaufman Assessment Battery for Children (KABC; Kaufman & Kaufman, 1983), McCarthy Scales of Children’s Abilities (MSCA; McCarthy, 1972), Differential Ability Scales (DAS; Elliot, 1990), Stanford-Binet Intelligence Scales – Fourth Edition (SB–IV; Thorndike, Hagen & Sattler, 1986), and Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993).

Specifically, Whitaker (2010) discussed the greater increases in novel tasks versus crystallized tasks as noted on the KABC, with greater gains on the Simultaneous Processing scale than the Achievement scale, greater gains on the Abstract/Visual

Reasoning scale than the Verbal Reasoning scale of the SB–IV, and greater gains on the Fluid IQ than the Crystallized IQ on the KAIT.

Theoretical perspectives on the effects of aging on cognition. The discussion of test-retest practice effects above focused on studies involving children's performance. Because the current study is using data from the WAIS–III, an adult assessment of intellectual functioning, it is necessary to review the effects of aging on cognition to understand how test-retest results may or may not be different for adults and children. A review of the literature indicates the existence of several classical theories of aging (Kaufman & Lichtenberger, 2006). Botwinick (1977) concluded that as we age, our ability to perform timed tasks becomes compromised, while our performance on nontimed tasks is left intact and spared. This theory is relatively consistent with the crystallized/fluid theory of cognitive development first posited by Horn and Cattell (1966, 1967). According to Kaufman and Lichtenberger (2006), this theory states that fluid intelligence, characterized by novel problem solving and abstract reasoning, has a pattern of increasing with age until young adulthood, at which point it starts to decline as neurological decay begins to occur. Conversely, crystallized intelligence, which is characterized by codified knowledge gained through formal education and life experiences, remains spared as we age and has the potential to increase throughout the lifespan as we add to our store of knowledge and experience.

More recently, yet another theory of cognitive aging has been discussed. Baltes (1997) developed a theory of cognitive development that consists of two distinct components: pragmatics and mechanics. The concept of pragmatics is analogous to the concept of crystallized ability noted above and is believed to be spared the effects of cognitive degradation. Mechanics, conversely, are presumed to be adversely impacted as the brain ages and include aspects of cognition such as visual processing, processing speed, and aspects of memory (Kaufman & Lichtenberger, 2006).

To lend credence to the aforementioned theories of aging, Kaufman and Lichtenberger (2006) highlight multiple studies that have been completed. Cross-sectional studies found that when the amount of education is controlled, verbal abilities (roughly analogous to the aforementioned concept of crystallized ability) on measures such as the WAIS, WAIS–R, WAIS–III, KAIT, K–BIT, K–SNAP, and K–FAST had a consistent pattern of maintenance throughout the lifespan before a decline was noted as subjects reached their mid-70s and 80s. Conversely, performance abilities (roughly analogous to fluid ability) were found to be adversely impacted by age, with peak performance noted in the early 20s, followed by subsequent decline (Kaufman & Lichtenberger, 2006). Similarly, longitudinal studies utilizing cohorts from both the WAIS and WAIS–R standardization sample, as well as cohorts from the WAIS–R and WAIS–III normative sample, had the same patterns of aging as in the cross-sectional studies: maintenance of verbal abilities and significant degradation of performance abilities (Kaufman & Lichtenberger, 2006). Taken together, these studies lend support to the various theories of aging noted above.

Physical signs of brain aging: As the brain ages, there are certain physiological changes that occur. Chief among these are a widening of the sulci and enlargement of the ventricles (Rettmann, Prince, & Resnick, 2003). Goldberg (2009) notes that these physiological changes are indicative of neural changes taking place within our brains, such as atrophy, the loss of the myelin sheath that insulates axons and speeds transmission, and reduction of dendritic branching. Further, Goldberg indicates that there are two general rules when it comes to the aging brain. The first is that the prefrontal cortex is especially vulnerable to normal age-related decay. Because the frontal lobes are crucial to helping us navigate novel situations, one of the adverse impacts of cognitive aging is that elderly individuals often have difficulty when presented with new situations and ingrained patterns are disturbed (Goldberg, 2009). Secondly, rates of age-related decay between the right and left hemisphere are not uniform, with the right hemisphere displaying loss of functioning due to aging earlier than the left hemisphere. Goldberg goes on to summarize various studies that have lent support to this pattern. Rettmann, Prince, and Resnick (2003) found that the sulci of aging brains became shallower (and thus decreased in functionality) in the right occipital and parietal regions than in similar cortical structures of the left hemisphere. Similarly, Grieve, Clark, and Gordon (2003) noted age- related vulnerability of the right hemisphere in comparison to the left hemisphere. They specifically found that age-related decay of the insula, a region implicated in integration of sensory information, was greater in the right hemisphere than in the left. A study conducted in Japan utilized magnetic resonance imaging (MRI) technology to show that age-related decline in brain region size occurs earlier in the right hemisphere than it does in the left hemisphere. The study specifically noted that this loss of brain volume begins in the fourth decade of life and intensifies during the fifth decade. Conversely, volume loss in the left hemispheric does not begin until the fifth decade (Taki et al., 2003).

It should also be noted that these rates of hemispheric differences in degradation are not limited to the cortex (Goldberg, 2009). Another MRI study found that in a population of elderly individuals with depression, the frontal lobes and right hippocampus had a loss of volume whereas the left hippocampus did not have the same loss (Ballmaier, Kumar, Elderkin-Thompson, 2003).

Perhaps most interesting, Goldberg (2009) goes on to assert that age-related decline in areas involved in novel problem solving, such as the frontal lobes and right hemisphere, and the preservation of the left hemisphere, the storehouse of established knowledge, may reflect an increased reliance on codified or crystallized knowledge and a decreased need for dealing with novelty as we age. Although it is beyond the scope of this study, these concepts represent important implications in areas such as ongoing neuroplasticity and whether humans have the ability to maintain aspects of cognitive functioning throughout the lifespan through practice and use.

Studies examining practice effects related to cognitive aging. Short-term practice effects in an elderly population were examined in a study completed by Krenk, Rasmussen, Siersma, and Kehlet (2012). They examined 161 healthy adults 60 of age or older on seven neuropsychological tests measuring aspects of set shifting, declarative memory, executive flexibility working memory, visuomotor skills, auditory recall, selective attention, interference susceptibility, and processing speed. Testing was performed on three occasions: at baseline, 1 week, and 3 months. Results indicated a statistically significant improvement in test scores in two of seven measures between baseline and the second administration and in six of seven measures between baseline and the final administration (Krenk et al., 2012).

Thorvaldsson, Hofer, Berg, and Johansson (2006) utilized practice effects data from a longitudinal study that started in 1971 on two different sets of participants. Data was obtained from subtest scores from the psychometric battery of Dureman and Salde (1959), such as Synonyms, Block Design, Figure Identification, Identical Forms, and Digit-Span Forward/Backward. One group was measured at ages 70, 75, 79, 81, 85, 88, 90, 92, 95, 97, and 99. The other group was measured at ages 85, 88, 90, 92, 95, 97, and 99. Results indicated a trend toward test effects on two of the five measures: the Synonyms and Block Design subtests (Thorvaldsson et al., 2006).

As noted above, classical test theory has typically eschewed the use of subtest level analysis due to poor reliability typically associated with subtests. However, it is apparent that different conceptualizations and calculation methods of reliability not entrenched in traditional psychometric theory provide valuable representations of test reliability. Viewed in this light, it is apparent that what traditionally has been viewed as error, causing instability of scores from time 1 to time 2, could actually be the result of the inadequacies of traditional reliability calculation methods (Whitaker, 2010). When subtests are examined using methods that reveal a greater degree of reliability than previously thought, deeper avenues of diagnostic interpretation and intervention planning can be gleaned from the subtests. This increases the overall clinical utility of intelligence tests.

Taking this a step further, Whitaker (2010) notes that the traditional psychometric theory approach to reliability is based on the assumption that individual test performance should not vary between time 1 and 2 because the cognitive capacities that these measures assess are thought to be stable over time. Further, when performance does increase from time 1 to time 2, classical test theory attributes such improvement to measurement error, rather than systematic variance from other sources (Whitaker, 2010).

Not only are there apparent methodological problems with the traditional psychometric calculations and conceptualizations of reliability, there are also shortcomings noted when reviewing the neuropsychological literature (Whitaker, 2010). It was specifically noted that viewing variations in performance from time 1 to time 2 as measurement error is not consistent with neuropsychological research that identifies changes in performance as the result of increased cognitive efficiency associated with practice effects (Catron & Thompson, 1979; Goldberg, 2009; Kaufman, 2003; Matarazzo Carmody, & Jacobs,, 1980).

Changes in brain functioning that enable once-novel tasks to become more routinized, and thus to be performed more effectively when repeated, are hypothesized to be the cause of score increases from time 1 to time 2 (Whitaker, 2010). Thus, what has been typically associated with random error is actually the result of physiological brain changes. This makes conceptual, neuropsychological, and common sense. Perhaps most importantly, it leads to increased clinical diagnostic utility of measurements.

Decision-consistency models similar to that proposed by McCloskey (1990) can provide a starting point for further development of alternative methods of reliability estimation that are free of the shortcomings of the traditional psychometric approach, while at the same time being more consistent with the neuropsychological literature (Whitaker, 2010).

Chapter 3

Method

This study utilized the archival data set from the WAIS–III standardization sample used to calculate the test-retest reliability for this instrument. As reported in the *WAIS–III Technical and Interpretative Manual* (Psychological Corporation, 1997), the participants included 394 adults (52.3% female and 47.7% male) between the ages of 16 and 89 years selected to be representative of the total WAIS–III standardization sample. Each participant was assessed on two occasions (mean test-retest interval, 32 days) using all 15 subtests of the WAIS–III. Other demographic characteristics of the sample are as follows: 74.1% Caucasian, 7.8% African American, 11.1% Hispanic, and 7% other racial/ethnic origin. Parent education levels for the participants were: 0-8 years, 4.9%; 9-11 years, 9.1% ; 12 years, 25.9%; 13-15 years, 36.2%; and 16 years or older, 23.9% (Psychological Corporation, 2004).

Measures.

The WAIS–III yields standard and scaled scores, base rates, percentile ranks, and age equivalents. Subtests have a mean scaled score of 10 and standard deviation of 3. The mean standard score for the four indexes is 100, and the standard deviation is 15. The *WAIS–III Technical and Interpretative Manual* (Psychological Corporation, 1997) provides detailed information regarding the instrument’s validity and reliability. The WAIS–III has been shown to have adequate content, criterion-related, and construct validities. As reported in the *WAIS–III Integrated Technical and Interpretative Manual* (Psychological Corporation, 1997), Tables 2 and 3 show the average standard error of the mean, reliability coefficients, and corrected stability coefficients for the subtests, composite scales, and process scores for the total sample.

Table 2

Average Reliability Coefficients, Standard Error of the Mean (SEM), *and Corrected Stability Coefficients for WAIS–III Subtests for the Total Sample*

Subtest	r_{xx}	*SEM*	*r*
Picture Completion	.83	1.25	.82
Vocabulary	.93	0.79	.93
Digit Symbol-Coding	.84	1.19	.82
Similarities	.86	1.12	.84
Block Design	.86	1.14	.89
Arithmetic	.88	1.05	.88
Matrix Reasoning	.90	0.97	.89
Digit Span	.90	0.94	.91
Information	.91	0.91	.92
Picture Arrangement	.74	1.53	.72
Comprehension	.84	1.21	.85
Symbol Search	.77	1.43	.77
Letter–Number Sequencing	.82	1.30	.79
Object Assembly	.70	1.66	.73

Note. r_{xx} = overall average reliability coefficient; *r* = stability coefficient. Overall average reliability and stability coefficients were calculated using the formula for Fisher's *z* transformation recommended by Silver and Dunlap (1987). Stability correlations were corrected for variability of the standardization sample using the procedures recommended by Allen and Yen (1979) and Magnusson (1967). Average *SEM*s were calculated by averaging the sum of the squared *SEM*s for each age group and obtaining the square root of the result.

Table 3

Average Reliability Coefficients, Standard Error of the Mean (SEM)*, and Corrected Stability Coefficients for WAIS–III Composite Scales for the Total Sample*

Process Score	r_{xx}	*SEM*	r^a
Verbal Comprehension	.96	3.01	.96
Perceptual Organization	.93	3.95	.94
Working Memory	.94	3.84	.93
Processing Speed	.88	5.13	.87
Full Scale IQ	.98	2.30	.98

Note. r_{xx} = overall average reliability coefficient; r = stability coefficient. Overall average reliability and stability coefficients were calculated using the formula for Fisher's *z* transformation recommended by Silver and Dunlap (1987). Stability correlations were corrected for variability of the standardization sample using the procedures recommended by Allen and Yen (1979) and Magnusson (1967). Average *SEM*s were calculated by averaging the sum of the squared *SEM*s for each age group and obtaining the square root of the result. Internal consistency coefficients for the indexes ranged from .88 for the PSI to .97 for the FSIQ. However, it is important to note that the internal reliability estimates for the PSI subtests are actually the test-retest reliability estimates; internal consistency estimates are not calculated for processing-speed subtests. Symbol Search and Cancellation had the lowest internal reliability estimates (.79), while Letter-Number Sequencing had the highest (.90). Internal consistency estimates for the process scores range from .70 for Cancellation Random to .84 for Block Design No Time Bonus.

The average standard error of the mean *(SEM)* is lowest for Vocabulary (.79) and highest for Object Assemble (1.66). For the composite scales and process scores, the *SEM* is lowest for FSIQ (2.30) and highest for Perceptual Organization Index (3.95). Composite test-retest coefficients for the total sample ranged from .98 for the FSIQ and to .88 for the PSI. Vocabulary had the highest test-retest reliability estimate (.93), and Object Assembly had the lowest (.70).

Research design and statistical procedures.

A modified decision-consistency model was used to categorize test-retest results by degree of change from time 1 to time 2. Using the time 1 and time 2 test scores of 394 cases from the standardization test-retest reliability study, the following procedures were conducted:

1. For each WAIS–III subtest score, an actual difference score was calculated for each case by subtracting the obtained time 1 score from the obtained time 2 score.
2. For each WAIS–III subtest score, a predicted time 2 score was calculated using the following formula:

   x2 = X2 + (x1 – X1) * *r*x1x2 * (*SD*x2/*SD*x1), wherex2 = time 2 score

   x1 = time 1 score

   X1 = mean of time 1 scores (set at 10 for all subtests)

   *r*x1x2 = the correlation between time 1 and time 2 scores

   *SD*1 = the standard deviation of time 1 scores (set at 3)

   *SD*2 = the standard deviation of time 2 scores (set at 3)

   It is important to note that the formula for predicting a time 2 score from a time 1 score is identical to the formula for estimating the true score because the term

(*SD*2/*SD*1) is equal to 1; both standard deviations were set at the known population value of 3. It is also important to note that the formula incorporates the concept of regression to the mean in that time 1 scores well below the mean of 10 are predicted to increase toward the mean, whereas time 1 scores well above the mean of 10 are predicted to decrease toward the mean.

3. For each WAIS–III subtest score, a predicted difference score was calculated for each case by subtracting the obtained time 1 score from the predicted time 2 score calculated in step 2.
4. For each WAIS–III subtest, frequency distributions were obtained for the actual difference scores and the predicted difference scores and placed together in a table for comparison and analysis.
5. For each WAIS–III subtest, the subtest scores of the sample of 394 cases were divided into two groups to test the traditional psychometric conceptualization of regression to the mean that was incorporated in the formula for predicted time 2 score.
    a. The LTE (less than or equal to the mean scaled score of 10) group consisted of all cases in which the time 1 score was less than 10. This group consisted of the cases predicted to show no change in score at time 2 or, in extreme cases, to show a positive gain in score at time 2 due to regression to the mean. Also included with this group were cases in which the time 1 score was 10 and the time 2 score was less than 10 for reasons stated in item c below.
    b. The GTE (greater than or equal to the mean scaled score of 10) group consisted of all cases in which the time 1 score was greater than 10. This group consisted of

the cases predicted to show no change in score at time 2 or, in extreme cases, to show a decrease in score at time 2 due to regression to the mean. Also included in this group were cases in which the time 1 score was 10 and the time 2 score was greater than 10 for reasons stated in item c below.

c. Cases scoring at the mean of 10 at time 1 presented a challenge in terms of group classification. Because all of these cases were at the mean at time 1, they were predicted to remain at the mean at time 2. If the predicted difference score of 0 was not identical to the actual difference score, then these cases would not be conforming to the expected pattern of regression to the mean. Because the analysis was attempting to determine the number of cases that did not conform to the expected pattern of regression to the mean, it was decided to maintain these cases in the analysis by dividing them based on the actual difference score to reflect their lack of conformity to the expected pattern. Cases earning time 1 scores of 10 that reflected a negative actual difference score (reflecting time 2 movement away from the mean instead of to the mean) were included in the LTE group, and cases earning a time 1 score of 10 that reflected a positive actual difference score (reflecting time 2 movement away from the mean instead of to the mean) were included in the GTE group.

   Time 1 scores of 10 that corresponded to Time 2 scores of 10 were consistent with the traditional psychometric model that would predict that scores at the mean would remain at the mean. These cases were divided evenly between the LTE and GTE groups, thereby adding equally to the number of cases in each group that conformed to the traditional psychometric model.

6. For each WAIS–III subtest, predicted difference scores were assigned to one of three score-change categories: negative change (-), no change (0), and positive change (+).
7. For each WAIS–III subtest, actual difference scores were assigned to one of three score-change categories: negative change (-), no change (0), and positive change (+).
8. For each WAIS–III subtest, a 2 x 5 cross-tabulation table for the cases assigned to the LTE group was generated indicating frequency counts of the score-change categories crossed with Time 1 scores less than 10 and Time 1 scores equal to 10 for actual difference scores and predicted difference scores as shown in Table 4. The actual (observed) and predicted (expected) frequencies in the 2 x 5 table were subjected to a chi-square analysis to determine goodness of fit between the actual and the predicted difference proportions.

Table 4

Cross-tabulation for cases assigned to the LTE group.

	Time 1 Score < 10			Time 1 Score = 10	
Actual Difference	Time 2 Score < Time 1 Score	Time 2 Score = Time 1 Score	Time 2 Score < Time 1 Score	Time 2 Score < Time 1 Score	Time 2 Score = Time 1 Score (1/2 of cases)
Predicted Difference	Time 2 Score < Time 1 Score	Time 2 Score = Time 1 Score	Time 2 Score < Time 1 Score	Time 2 Score < Time 1 Score	Time 2 Score = Time 1 Score (1/2 of cases)

Note. LTE = less than or equal to the mean scaled score of 10.

9. For each WAIS–III subtest, a 2 x 5 cross-tabulation table for the cases assigned to the GTE group was generated, indicating frequency counts of the score-change categories crossed with time 1 scores less than 10 and time 1 scores equal to 10 for actual difference scores and predicted difference scores, as shown in Table 5. The actual (observed) and predicted (expected) frequencies in the 2 x 5 table were subjected to a chi-square analysis to determine goodness of fit between the actual and the predicted difference proportions.

Table 5

Cross-tabulation for cases assigned to the GTE group.

	Time 1 Score < 10			Time 1 Score = 10	
Actual Difference	Time 2 Score < Time 1 Score	Time 2 Score = Time 1 Score	Time 2 Score < Time 1 Score	Time 2 Score < Time 1 Score	Time 2 Score = Time 1 Score (1/2 of the cases)
Predicted Difference	Time 2 Score < Time 1 Score	Time 2 Score = Time 1 Score	Time 2 Score < Time 1 Score	Time 2 Score < Time 1 Score	Time 2 Score = Time 1 Score (1/2 of the cases)

10. In many instances, the frequency counts for predicted difference scores were 0 for the negative-change and positive-change categories. Chi-square analyses require a minimum of five cases in each category. When the score-change category count was 0 for the predicted difference score, up to five cases were removed from the no change predicted category and placed in the category with the 0 count, thereby enabling the completion of all chi-square analyses. This alteration of the data represents a bias in favor of a nonsignificant finding in that increasing the category frequency for cells with 0 counts made it more likely that the proportions in each

category would be similar and more likely to result in a nonsignificant chi-square value.

Chapter 4

Results

Cross-tabulation analyses were conducted to determine the frequency of score differences between time 2 and time 1 administration for each WAIS–III subtest. The actual difference scores were then compared with the predicted difference scores calculated using the regression model described in Chapter 2.

Actual and predicted score frequency distributions.

Table 6 shows the frequency distributions for the actual and predicted differences between time 2 and time 1 performance on the WAIS–III Verbal subtests. For all Verbal subtests, with the exception of only one case for the Similarities subtest, the predicted difference scores did not exceed -1 or +1. In terms of the actual differences, across all Verbal subtests, a higher frequency of examinees had no change or positive scaled score change versus negative scaled score change. For all subtests, the retest performance of a majority of examinees was between -1 and 1, suggesting that this score band may prove useful in predicting WAIS–III Verbal retest performance using the alternative reliability model presented here. A fairly large discrepancy between actual and predicted score frequencies was evident for all subtests. In addition, when viewing the results in the table, a majority of no change or positive increases rather than decreases is apparent. That is, there is a preponderance of actual difference increases that are more consistent with a neuropsychologically based performance model, which takes into account brain-based functioning, rather than with the traditional psychometric model, which attributes variations in performance to random error.

Table 6

Frequency Distributions of Actual and Predicted T2-T1 Differences for Each WAIS–III Verbal Subtest

	Difference														
Vocabulary	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
(*n* = 394)															
Actual				3	5	16	74	148	101	35	8	2		1	1
Predicted								394	1						
Information	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
(*n* = 394)															
Actual					1	4	42	155	115	56	19	2			
Predicted							9	381	4						
Similarities	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
(*n* =394)															
Actual		1		5	9	24	61	119	76	54	24	12	6		2
Predicted						1	43	309	41						
Comprehension	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
(*n* = 394)															
Actual			1	5	20	33	79	90	82	40	27	11	2	3	1
Predicted							48	294	52						

The greatest magnitude of frequency of score variability from time 1 to time 2 was observed in the Similarities subtest, with two examinees having score improvements of +7 and 9 and one subject having a decrease of -6 scaled score points. In contrast to Similarities, analysis of Vocabulary and Comprehension subtest results from time 1 to time 2 yielded similar test-retest score frequency variability. Although still revealing the increases in performance evident in the other verbal subtests, the range of variability in scores from time 1 to time 2 for Information was slightly less, with two examinees having scaled score point gains of +4 and one examinee having a decrease in performance of -3 points.

Table 7 shows the frequency distributions for the actual and predicted differences between time 2 and time 1 performance on the WAIS–III Working Memory subtests. With the exception of Letter-Number Sequencing, which yielded one case with a gain of 2 scaled score points, remaining Working Memory subtest scores did not exceed -1 or +1 from time 1 to time 2 when utilizing the predicted difference method espoused by the traditional psychometric approach. However, just as with the Verbal subtests, fairly large discrepancies between actual and predicted score frequencies were evident for all subtests, although these discrepancies were somewhat smaller for the Letter-Number Sequencing Subtest. The increase, rather than the decrease in performance evident in the Verbal subtests, was again found here. That is, a greater number of examinees had no change or a positive scaled score change on the Working Memory subtests. Stated another way, performance on all Working Memory subtests was characterized by much higher frequencies of actual difference increases than actual difference decreases, a pattern that is more +consistent with a neuropsychologically based performance model than with the traditional psychometric model.

Table 7

Frequency Distributions of Actual and Predicted T2-T1 Differences for Each WAIS–III Working Memory Subtest and Process Score

							Difference								
Digit Span (*n* = 394)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual		1	2	4	11	31	89	83	74	56	24	11	5	3	
Predicted							9	377	8						
Arithmetic (*n* = 393)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual			1	1	12	29	68	109	72	61	31	4	6		
Predicted							34	320	40						
Letter-Number Sequencing (*n* = 374)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual		3	7	7	20	33	57	70	67	49	27	16	12	4	2
Predicted							75	225	73	1					

Table 8 shows the frequency distributions for the actual and predicted differences between time 2 and time 1 performance on the WAIS–III Perceptual Reasoning subtests. For Perceptual Reasoning subtests, the predicted difference frequency scores from time 1 to time 2 were somewhat more variable. Predicted difference frequency scores for Block Design and Matrix Reasoning did not exceed -1 or +1. Similarly, predicted difference scores on Picture Completion ranged from -2 to +1. Range of frequencies for predicted differences for Picture Arrangement and Object Assembly were slightly higher, with predicted difference scores ranging from -3 to +2 and -3 to +3, respectively. However, just as with the Verbal Comprehension and Working Memory subtests, predicted scores based on the traditional psychometric model had only limited predictive validity in comparison to actual performance differences between time 1 and 2. Although score differences for the Picture Arrangement and Object Assembly Subtests showed greater variability for the predicted scores, even the predicted score ranges for these subtests were very restricted when compared to the ranges of the actual scores. As was evident for the Verbal Comprehension and Working Memory subtests, a much greater number of examinees had no change or positive scaled score changes rather than negative changes across all Perceptual Reasoning subtests. For all subtests, large numbers of examinees had no scaled score change between time 1 and time 2. Block Design, Matrix Reasoning, Picture Arrangement, and Object Assembly yielded the largest range of actual score variability from time 1 to time 2, with examinee performance ranging from -7 to +9. Picture Arrangement, Picture Completion, and Object Assembly had the highest number of examinees with score improvements both within the Perceptual Reasoning Index and

Table 8

Frequency Distributions of Actual and Predicted T2-T1 Differences for Each WAIS–III Perceptual Reasoning Subtest

	Difference														
Block Design (*n* = 394)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual	1			4	13	36	52	91	73	71	37	12	1	2	1
Predicted							44	306	44						
Matrix Reasoning (*n* = 394)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual	1	1	4	9	21	42	58	95	73	40	30	10	7	1	2
Predicted							14	365	15						
Picture Completion (*n* = 394)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual					7	11	23	73	71	62	65	36	22	19	5
Predicted						1	67	267	59						

(continued)

	Difference														
Picture Arrangement (*n* = 394)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual	1	2	4	8	11	27	38	78	59	58	47	20	19	11	11
Predicted					1	16	61	216	92	8					
Object Assembly (*n* = 393)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual	1		1	5	10	16	38	78	66	59	45	27	27	11	11
Predicted					1	17	82	175	100	18	1				

the WAIS–III as a whole, with Picture Completion having the largest number of the three subtests. Overall results indicate that the actual performance increases observed are much more consistent with a neuropsychologically based performance model than a traditional psychometric model involving regression to the mean.

Table 9 shows the frequency distributions for the actual and predicted differences between time 2 and time 1 performance for the WAIS–III Processing Speed subtests and process scores. The predicted difference score ranges for Digit Symbol-Coding and Symbol Search were -1 to +1 and -2 to +2, respectively. However, most importantly, for both subtests there was a large discrepancy between the frequencies predicted by the regression model based on a traditional psychometric approach and the actual differences that occurred. As with the Verbal Comprehension, Working Memory, and Perceptual Reasoning subtests, a greater number of examinees had no scaled score change or positive scaled score changes, rather than negative scaled score changes, for both Processing Speed Subtests. Thus, as with the subtests in the other domains, performance on the Processing Speed subtests was characterized by higher frequencies of actual difference increases, rather than decreases that are more consistent with a neuropsychologically based performance model.

Table 9

Frequency Distributions of Actual and Predicted T2-T1 Differences for Each WAIS–III Processing Speed Subtest

	Difference														
Digit Symbol-Coding (*n* = 393)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual			1	4	6	13	37	88	113	71	41	10	5	2	2
Predicted							61	274	59						
Symbol Search (*n* = 394)	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual	1		2	7	8	31	50	108	90	49	27	13	4	1	3
Predicted						5	81	223	78	7					

Comparison of actual and predicted score differences according to time 1 score groups.

Tables 5 through 9 present actual and predicted score frequency data for the total sample without indication of the scores obtained at time 1. To more clearly examine the effectiveness of the psychometric model in predicting the time 2 score, it is necessary to specify the value of the time 1 score to examine the expected effects of regression to the mean. Per traditional psychometric conceptualizations, the regression model predicts time 2 score increases for more extreme time 1 scores below the mean and time 2 score decreases for more extreme time 1 scores above the mean.

Tables 10 through 13 show the percentages of time 2 to time 1 score differences that reflect decreases (T2 < T1), increases (T2 > T1), or no change (T2 = T1) separately for examinees divided into two categories: examinees earning scores less than or equal to the mean scaled score of 10 (LTE group) and examinees earning scores greater than or equal to the mean scaled score of 10 (GTE group). Also provided are the percentages of time 2 to time 1 score differences that the regression model predicts would result in score decreases, increases, or no change for the examinees in the LTE and GTE groups. Examinees scoring at the mean were divided equally between the LTE and GTE groups, as described in Chapter 3.

Table 10

Actual and Predicted Difference Score Percentages by Score Change Category for Each WAIS–III Verbal Subtest for the Less Than or Equal to (LTE) and Greater Than or Equal to (GTE) Groups

	T1 Scaled Scores			T1 Scaled Scores		
	LTE Group			GTE Group		
	T2 < T1	T2 = T1	T2 < T1	T2 < T1	T2 = T1	T2 < T1
	Vocabulary (*n* = 190)			Vocabulary (*n* = 204)		
Predicted	0%	99%	1%	0%	100%	0%
Actual	24%	42%	34%	25%	34%	41%
	Information (*n* = 176)			Information (*n* = 218)		
Predicted	0%	98%	2%	4%	96%	0%
Actual	11%	46%	43%	12%	34%	54%
	Similarities (*n* = 181)			Similarities (*n* = 213)		
Predicted	0%	77%	23%	21%	79%	0%
Actual	19%	36%	45%	31%	25%	44%
	Comprehension (*n* = 177)			Comprehension (*n* = 177)		
Predicted	0%	71%	29%	22%	78%	0%
Actual	29%	24%	47%	40%	22%	38%

Note. LTE = time 1 standard score below the mean of 10. GTE = time 1 standard score above the mean of 10. Time 1 scores at the mean were divided evenly between the two groups.

For all but one of the Verbal Comprehension subtests, both the LTE and GTE groups had higher percentages of actual score increases than decreases. The only exception to this pattern was on the Comprehension subtest, which yielded a slightly greater percentage of decreases than increases(40% vs. 38%). Also, all Verbal Comprehension subtests had lower percentages of no change than predicted. Score increase percentages for the LTE group in comparison to the GTE group were somewhat variable, with higher score increase percentages for the GTE group noted on the Vocabulary (41% vs. 34%) and Information (54% vs. 43%) subtests and higher LTE progression percentages on the Similarities (45% vs. 44%) and Comprehension (47% vs. 38%) subtests. Across all subtests, both groups had higher percentages of actual score increases than predicted. Counter to the expectations of a traditional psychometric model, the GTE group had high rates of actual versus predicted score increases for all Verbal Comprehension subtests. Also counter to the traditional psychometric model, the LTE group had greater numbers of examinees earning both increased and decreased scores.

Table 11

Actual and Predicted Difference Score Percentages by Score Change Category for Each WAIS–III Working Memory Subtest and Process Score for the Less Than or Equal to (LTE) and Greater Than or Equal to (GTE) Groups

	T1 Scaled Scores			T1 Scaled Scores		
	LTE Group			GTE Group		
	T2 < T1	T2 = T1	T2 < T1	T2 < T1	T2 = T1	T2 < T1
	Digit Span (*n* = 230)			Digit Span (*n* = 164)		
Predicted	0%	97%	3%	5%	95%	0%
Actual	35%	20%	45%	35%	23%	42%
	Arithmetic (*n* = 186)			Arithmetic (*n* = 208)		
Predicted	0%	78%	22%	16%	84%	0%
Actual	28%	24%	48%	28%	31%	41%
	Letter-Number Sequencing (*n* = 175)			Letter-Number Sequencing (*n* = 199)		
Predicted	0%	58%	42%	38%	62%	0%
Actual	29%	20%	51%	38%	18%	44%

Note. LTE = time 1 standard score below the mean of 10. GTE = time 1 standard score above the mean of 10. Time 1 scores at the mean were divided evenly between the two groups.

For all Working Memory subtests, both the LTE and GTE groups had higher percentages of actual score increases versus actual score decreases, as well as lower percentages of no change scores than predicted. The percentages of increased scores in the LTE group were higher than in the GTE group. Across all subtests, both groups had higher percentages of actual score increases than predicted. The LTE group had higher rates of actual versus predicted score decreases for all Working Memory subtests. The GTE group had higher rates of actual versus predicted score decreases for Digit Span, Arithmetic, and Letter-Number Sequencing (38% for each). Actual decreased score rates for the LTE and GTE groups were somewhat variable, with the same rate for Arithmetic (28% for each) and Digit Span (35% for each) and a higher percentage for the GTE group (38%) than the LTE group (29%) for Letter–Number Sequencing.

Table 12

Actual and Predicted Difference Score Percentages by Score Change Category for Each WAIS–III Perceptual Reasoning Subtest for the Less Than or Equal to (LTE) and Greater Than or Equal to (GTE) Groups

	T1 Scaled Scores			T1 Scaled Scores		
	LTE Group			GTE Group		
	T2 < T1	T2 = T1	T2 < T1	T2 < T1	T2 = T1	T2 < T1
	Block Design (*n* = 194)			Block Design (*n* = 200)		
Predicted	0%	77%	23%	22%	78%	0%
Actual	27%	19%	54%	26%	27%	47%
	Matrix Reasoning (*n* = 202)			Matrix Reasoning (*n* = 192)		
Predicted	0%	93%	7%	7%	93%	0%
Actual	30%	27%	43%	39%	21%	40%
	Picture Completion (*n* = 171)			Picture Completion (*n* = 223)		
	0%	65%	35%	30%	70%	0%
	12%	15%	73%	9%	22%	69%
	Picture Arrangement (*n* = 187)			Picture Arrangement (*n* = 207)		
	0%	46%	54%	38%	62%	0%
	19%	20%	61%	26%	20%	54%
	Object Assembly (*n* = 186)			Object Assembly (*n* = 207)		
Predicted	0%	36%	64%	48%	52%	0%
Actual	18%	20%	62%	18%	19%	63%

Note. LTE = time 1 standard score below the mean of 10. GTE = time 1 standard score above the mean of 10. Time 1 scores at the mean were divided evenly between the two groups.

Both the LTE and GTE groups had higher percentages of actual score increases than decreases for all subtests. Picture Completion showed the highest percentages of score increases, with 73% of the LTE group and 69% of the GTE group having higher scores at time 2. In addition, for both groups, the predicted percentages of no score change were much higher than the percentages of actual score changes. For the majority of Perceptual Reasoning subtests, the LTE group had higher percentages of actual score increases than the GTE group. The exception was the Object Assembly subtest, in which the GTE group had a higher percentage of actual score increases than the LTE group (63% vs. 62%). Comparison of actual decreased score percentages between the LTE and GTE groups yielded some variations, depending on the subtest. Actual increased score percentages were greater for the GTE group for the Picture Arrangement and Matrix Reasoning subtests. Conversely, actual decreased score percentages were greater for the LTE group for the Picture Completion and Block Design subtests. The same decreased score percentage (18%) was found on the Object Assembly for both groups. For the majority of Perceptual Reasoning subtests across both groups, the predicted percentages of score increases were lower than the actual percentages for all subtests for both groups. The exception was the LTE group for the Object Assembly subtest, which had a higher predicted percentage than the actual percentage (64% vs. 62%). The LTE group had higher percentages of actual than predicted time 1 scaled scores. Conversely, the GTE group had more variation in percentages of actual versus predicted score decrease. Specifically, within the GTE group, lower percentages than predicted were found on the Object Assembly, Picture Arrangement, and Picture Completion subtests. Higher than predicted score decrease percentages were identified on the Matrix Reasoning and Block Design subtests.

Table 13

Actual and Predicted Difference Score Percentages by Score Change Category for Each WAIS–III Processing Speed Subtest for the Less Than or Equal to (LTE) and Greater Than or Equal to (GTE) Groups

	T1 Scaled Scores			T1 Scaled Scores		
	LTE Group			GTE Group		
	T2 < T1	T2 = T1	T2 < T1	T2 < T1	T2 = T1	T2 < T1
	Digit Symbol-Coding (*n* = 191)			Digit Symbol- Coding (*n* = 203)		
Predicted	0%	69%	31%	30%	70%	0%
Actual	13%	25%	62%	18%	20%	62%
	Symbol Search (*n* = 181)			Symbol Search (*n* = 213)		
Predicted	0%	53%	47%	40%	60%	0%
Actual	23%	25%	52%	27%	29%	44%

Note. LTE = time 1 standard score below the mean of 10. GTE = time 1 standard score above the mean of 10. Time 1 scores at the mean were divided evenly between the two groups.

For both Processing Speed subtests, both the LTE and GTE groups had higher percentages of actual score increases than actual score decreases and lower percentages than predicted of no change. Actual percentages of increased and decreased scores were higher than predicted for the LTE group for both subtests. For the GTE group, actual decreased score percentages were lower than predicted for both Processing Speed subtests had, and actual increased score percentages were higher than predicted.

Percentages of actual score increases for both groups were the same for Digit Symbol-Coding (62%) and somewhat higher for the LTE group for Symbol Search. Percentages of actual score decreases were higher for the GTE group than for the LTE group for all subtests.

Chi-Square analyses comparing actual and predicted scores by time 1 score groups. Chi-square analyses were conducted to determine the goodness of fit between actual and predicted time 2 to time 1 score differences for all subtests scores. As indicated in Table 14, there were statistically significant results ($p < .01$) for all subtests in both the LTE and GTE groups.

Table 14

Chi-Square Analysis of Actual (Observed) and Predicted (Expected) T2-T1 Differences for Each WAIS–III Subtest Grouped by Time 1 Standard Score Range Groups

	T1 Scaled Scores LTE Group			T1 Scaled Scores GTE Group		
	X^2	*df*	Cramer's *V*	X^2	*df*	Cramer's *V*
Vocabulary	934.8	4	.55	1147.1	4	.64
Information	1032.6	4	.55	1708.3	4	.62
Similarities	181.8	4	.43	798.6	4	.56
Comprehension	239.5	4	.48	757.4	4	.57
Block Design	371.5	4	.59	804.0	4	.55
Matrix Reasoning	757.3	4	.64	910.1	4	.70
Picture Completion	156.0	4	.52	2983.5	4	.69
Picture Arrangement	110.2	4	.36	1246.5	4	.58
Object Assembly	95.5	4	.29	1939.5	4	.64
Digit Span	1947.4	4	.75	774.8	4	.69
Arithmetic	464.2	4	.55	802.3	4	.54
Letter Number Sequencing	207.6	4	.43	794.2	4	.57
Digit Symbol-Coding	153.8	4	.45	1619.1	4	.63
Symbol Search	152.2	4	.36	835.4	4	.49

Note. p < .01

Chapter 5

Discussion

The first aim of this study was to compare a neuropsychologically based performance consistency model with the traditional psychometric model for the purposes of representing WAIS–III subtest test-retest data. The second aim of the study was to determine if a neuropsychologically oriented performance consistency method offered any advantages over traditional psychometric methods in the type of information provided regarding WAIS–III subtest performance patterns.

WAIS–III subtest data analyses comparing the traditional psychometric model with a neuropsychologically based performance consistency model.

Results indicated that the traditional psychometric approach typically yielded predicted score difference frequencies that did not vary beyond the range of -3 to +3 points. Most predicted values were 0, indicating the expectation of no change in performance from time 1 to time 2. However, actual performance results produced a very different pattern of score difference frequencies, yielding both larger score band widths (-9 to +9) and much larger numbers of examinees distributed across this greater range of scores. Regardless of the score in relation to the mean, most subtests had much greater frequencies of positive versus negative score changes beyond the -3 to +3 score band.

The greatest magnitude of score differences (4 points and greater) favored positive scaled score differences. Taken as a whole, this information suggests that a neuropsychologically based performance model yields a progression effect wherein more examinees had increases than decreases in performance when exposed to the tasks a second time, and the frequency and size of performance gains often were more than 3 scaled score points. Given what is known concerning brain changes with subsequent exposure to the same task, these results are concordant with the neuropsychological literature.

The present study also grouped the examinees based on their time 1 scores in relation to the mean (aforementioned LTE and GTE groups). When grouping the data in this manner, similar trends in performance were observed for both groups. Comparing the actual and predicted score differences utilizing chi-square analyses indicated a poor goodness of fit between what would be predicted by the traditional psychometric model and the actual results obtained using the neuropsychological performance model. These findings are inconsistent with a traditional conceptualization of reliability, which would predict some degree of regression to the mean for WAIS–III subtest scores both above and below the mean. For the GTE group, the majority of examinees were predicted to earn a time 2 score identical to their time 1 score. Most of the remaining examinees' scores were expected to regress to the mean, with time 2 scores 1 point lower than their time 1 score. For the LTE group, the majority of examinees also were predicted to earn a time 2 score identical to their time 1 score, and most of the remaining examinees' scores were expected to regress to the mean, with time 2 scores that were 1 point higher than their time 1 score.

As noted above, for a majority of cases, the psychometric model predicted no change in performance from time 1 to time 2. Any predicted change reflected regression to the mean, based on initial subtest performance. For example, for time 1 scores above the mean (GTE group), time 2 scores were predicted to decrease to the mean.

Conversely, for time 1 scores below the mean, time 2 scores were predicted to increase to the mean (progression in the form of regression to the mean).

Analysis of actual score performance for the LTE group revealed that for all subtests within all four composites, a much higher percentage of examinees had score increases toward the mean. Traditionally, psychometric theory would consider these increases to be due to regression to the mean and identify the variability from time 1 to time 2 as the effects of chance or random error. However, with the exception of Object Assembly, a greater percentage of actual score increases than predicted score increases also was noted for the GTE group. This finding is in direct contradiction to the traditional psychometric theory explanation of score changes, in that traditional theory predicts that scores above the mean at time 1 will produce scores closer to the mean at time 2. This finding for the GTE group suggests that what has traditionally been conceptualized as regression to the mean for the LTE group is actually progression beyond expected performance levels based on changes in brain function. For both the GTE and the LTE groups and for all subtests of the WAIS–III, there was a higher percentage of actual score progression than no change among examinees for whom no change was predicted.

For examinees scoring above the mean at time 1, either no change or negative change was predicted as a result of regression to the mean. Comparison of time 1 to time 2 results indicated that the GTE group did have some regression to the mean. However, as was the case with the LTE group, there were greater percentages of actual score progression away from the mean than regression to the mean for nearly all subtests. The exception was for the Comprehension Subtest, in which the proportion of cases with regression to the mean was slightly higher than the proportion with progression away from the mean (40% vs. 38%). This suggests that what has traditionally been conceptualized as random error in variability at time 2 is more likely to represent real systematic changes in brain function, producing learning and resulting in progression beyond expected performance levels. Therefore, support for a neuropsychological performance model of interpretation is reflected in the fact that all subtests had both greater proportions of actual progression than no change, and greater percentages of actual than predicted upward progression.

The finding of larger proportions of score increases than score decreases at time 2 for examinees scoring above the mean at time 1 (GTE group) strongly suggests that some portion of the large number of score increases observed for examinees scoring below the mean at time 1 (LTE group) was due not to regression to the mean as would be predicted by psychometric theory, but rather to changes in brain function, resulting in real learning and task improvement. Although the actual proportion of such real change versus change due to random error cannot be ascertained at this time, it is likely that the proportion of examinees in the LTE group having increases due to real changes in cognitive functioning is not greater than the proportion of examinees in the GTE group having progression from the mean at time 2 and is possibly smaller than the GTE proportion because examinees in the LTE group are likely to be less efficient learners than examinees in the GTE group.

The data analyses conducted indicate that the performance consistency model proposed by McCloskey (1990) provided a more neuropsychologically accurate representation of actual WAIS–III subtest test-retest patterns than what would be predicted by the traditional psychometric model, which hypothesized little change from time 1 to 2 and that attributes actual change to random error. It is important to note that the performance consistency model provided a more accurate representation regardless of whether the data was grouped according to the total sample or in reference to performance above or below the mean at time 1 (LTE and GTE groups). Cross-tabulation analyses for the total combined sample revealed statistically significant differences between the obtained and predicted score difference frequencies for all performance levels across all subtests.

Clinical utility of a decision consistency performance model compared to the traditional psychometric model.

The second aim of the study was to examine the usefulness of the neuropsychologically based performance model in terms of WAIS–III subtest interpretive information for clinicians in comparison to traditional psychometric conceptualizations of reliability. Whitaker (2010) noted that traditional psychometric theory assumes that intelligence is a static trait that does not change over time, especially when the time interval is short, and that changes in performance that do occur from time 1 to time 2 are due to noninterpretable random error, as opposed to clinically meaningful, interpretable changes.

If these conceptualizations about test-retest performance were clinically accurate, one would expect to see little or no difference between time 1 and time 2 performances in a group such as the WAIS–III test-retest sample. However, as shown in the analyses of the WAIS–III data, increases in scores from time 1 to time 2 were observed much more frequently than score decreases, and these increases often occurred for examinees who would be predicted to have scores with regression to the mean. Analyses conducted using the neuropsychologically based performance consistency model revealed that all but one of the subtests had a progression effect that was much more frequent than a regression effect. By inference, based on the performance of examinees who scored above the mean at time 1, it can be hypothesized that the time 2 gains made by many of the examinees who scored below the mean at time 1 also reflected a progression effect rather than a regression effect, especially in cases in which the gains were quite large (3 or more scaled score points).

Utilizing the same neuropsychologically based performance consistency model applied in this study, Whitaker (2010) found strikingly similar data trends that reflected progression (rather than regression) effects for all WISC–IV subtest and processing scores. According to psychometric theory, if random measurement error was the cause for variations in performance from time 1 to time 2, then such error should distribute randomly. Such random distribution should theoretically yield equal amounts of score regression and progression, depending on the relation of time 1 performance to the mean. However, this was not observed in the present study or in Whitaker's (2010) analysis of the WISC–IV data. When examining actual time 1 and time 2 performance, not only were there differences in comparison to predicted results, there also was a very strong overall progression effect both for examinees scoring below the mean at time 1 and for examinees scoring above the mean at time 1. These results are counter to the concept of random distribution of error that should have seen observed, based on psychometric theory.

The alternative model used here, first presented by McCloskey (1990) and later applied by Whitaker (2010), is a neuropsychologically based performance model that attempts to improve upon the traditional psychometric approach. The performance model appears to offer an improvement over the traditional psychometric model in two important and interconnected ways: it provides a better fit with what is known about brain function from a neuropsychological perspective, and it enables clinicians to derive greater clinical utility from their assessments. Perhaps most importantly, it takes into account the realities of brain-based behavior that occurs during engagement in cognitive tasks (Whitaker, 2010).

Based on a neuropsychological model, variations in test performance from time 1 to time 2 are not viewed as the result of random error, but rather as the result of real changes in brain function due to exposure and the nature of the cognitive task that is being performed. Specifically, the progression effect is hypothesized to result from greater efficiency in the expression of already learned skills, as well as the learning that occurs after the first exposure to a novel task (Whitaker, 2010). In a neuropsychological model, regression, when it occurs, is not conceptualized as random error, but rather as the result of less efficient application of cognitive capacities. Stated another way, progression and regression could be conceptualized as the degree of efficiency in the use of executive functions, as these are mental processes that cue the effective use of other cognitive capacities (McCloskey, Perkins, & Van Divner, 2009). Another characteristic of this theory is that not only would variations in test-retest performance occur, but these variations would not distribute randomly. That is, the progression effect typically will be larger than the regression effect for any group of examinees because the human brain changes when it is exposed to information, and the relative strength of this progression effect is a function of the cognitive processing demands of the tasks being performed.

The progression effects hypothesized based on a neuropsychological model were found in both the current study and the study conducted by Whitaker (2010). Whitaker (2010) noted that a review of the neuropsychological and education literature identified several cognitively based factors that will impact performance on subsequent administrations of a task, including psychomotor speed and the novelty of a given task. These factors reflect the cognitive processing demands necessary for successful task completion.

When discussing reliability, the results of test-retest analyses using the traditional psychometric model are typically presented in the form of a table listing the correlations between time 1 and time 2 performance. These correlations are interpreted as stability coefficients, representing the degree to which performance remains stable from time 1 to time 2. In this model, the variability present in scores at time 1 should remain the same at time 2, thereby producing a high correlation between performance at time 1 and time 2. The higher the correlation, the more performance at time 2 resembled exactly the performance at time 1 and the more stable the subtest performance. This presence of a high degree of stability reflected in a high correlation coefficient value is then interpreted as an indication of good reliability. Conversely, if many, but not all, scores increase from time 1 to time 2 and the increases are not consistent for all examinees, then the correlation coefficient will be much lower. The presence of a high degree of instability reflected in a low correlation coefficient value is then interpreted as an indication of poor reliability.

In the neuropsychological model, changes in performance from time 1 to time 2 that vary in magnitude from one examinee to another are not viewed as undesirable consequences of random error factors, but rather as expected outcomes, based on what is known about the changes that occur in human brains when these brains are exposed to information. Correlation coefficients based on stability of scores from time 1 to time 2 therefore are not likely to be the best way to understand or interpret the reliability of what is occurring with performance from time 1 to time 2. Rather, a table (such as Table 14) that presents the percentages of the total sample with positive change, negative change, and no change in test-retest scaled scores, along with the test-retest reliability coefficients derived from analysis of the data, may be a more effective way to help clinicians understand what to expect in terms of typical variations in examinee performance from time 1 to time 2.

Table 15 shows that for every subtest, the number of examinees with scaled score increases of 1 or more points was greater than the number who earned the same score and/or the number with scaled score losses of 1 or more points. It is equally important to note that the number of examinees with score decreases was always smaller than the number of examinees whose scores remained the same combined with the number of examines whose scores increased. When viewing the data, it also is clear that although more positive than negative change was noted overall, the magnitude of this change varied across the subtests. The key interpretive consideration here is that a neuropsychological model posits that progression is impacted by the cognitive demands of specific tasks.

Table 15

Summary of Percentage of Cases Within Score-Change Categories and Reliability Coefficients for Each WAIS–III Subtest and Selected Process Scores

Subtest	% Negative Change	% No Change	% Positive Change	Positive:Negative Change Ratio	r_{x1x2}
Vocabulary	25	37	38	1:52	.93
Information	12	39	49	4:08	.92
Similarities	25	30	45	1:80	.84
Comprehension	35	23	42	1:20	.85
Block Design	27	23	50	1:85	.89
Matrix Reasoning	35	24	41	1:17	.89
Picture Completion	10	19	71	7:10	.82
Picture Arrangement	23	20	57	2:48	.72
Object Assembly	18	20	62	3:44	.73
Digit Span	35	21	44	1:26	.91
Arithmetic	28	28	44	1:57	.88
Letter-Number Sequence	34	19	47	1:38	.79
Digit Symbol-Coding	16	22	62	3:88	.82
Symbol Search	25	27	48	1:92	.77

For example, given what has already been stated about how brains become more efficient at completing a novel task when exposed to it a second time, it is not surprising that the greatest percentages of change were noted with some of the Perceptual Reasoning and Processing Speed subtests. Subtests in these areas are the most novel on the WAIS–III and likely represent specific tasks that examinees had never been asked to perform. However, once the brain has been exposed to a particular task, the neuropsychological literature indicates that cognitive functioning shifts from right anterior sections of the cerebral cortex to left posterior sections as the brain classifies and codifies what it has learned and stores strategies for performing such tasks when encountered again. Once a task has been encountered, it is no longer truly novel, and the brain develops new and often more efficient strategies for problem solving when exposed to the task a second time. Gains in performance therefore could be conceptualized as the brain expressing the development and use of more efficient strategies.

These patterns in cognitive processing demands can be seen both across and within composites. For example, the data indicates that the most robust gains were made with the Picture Arrangement, Picture Completion, Object Assembly, and Digit Symbol-Coding subtests. These are the most novel of the subtests on the WAIS–III. In general, these are subtests in which the first exposure could be thought of as a learning trial in which the examinee is exposed to a novel task for which he or she must develop problem solving or reasoning strategies. During subsequent exposure to the tasks, the person does not completely relearn the task as if it were never seen before, but rather draws on the prior experience with the task to complete it in a more efficient manner. Whitaker (2010) described performance of the Picture Completion subtest to illustrate this. During Picture

Completion, the examinee is given 20 seconds to scan a picture and identify what essential aspect of the picture is missing. Progression gains on the second exposure to the task likely result from having already been exposed to the task, knowing what to look for and noticing details that may have been missed on the initial scan.

An analogy seems appropriate here. The first time that a person is driving to a new and unknown location would likely be characterized by periods of trial and error as he or she attempted to find their way. When asked to drive again to the same location, the person's performance would likely progress/improve as he or she has developed strategies/knowledge for how to find the location in a more efficient manner. If finding the unknown location were a formal test and repeated several times, the psychometric theoretical framework would espouse that the person may struggle just as much the next time they attempt to find the location, despite now having more knowledge and experience with the task and that their attempts would distribute fairly evenly around a true score. What seems more appropriate is that the person would draw on prior knowledge and problem solving strategies and become more efficient at reaching the destination each time he or she attempted to do so. Would not brains be expected to act in the same manner with subsequent exposure to most cognitive tasks? Does it not make more clinical, theoretical, practical sense and, based on the results of this study and that conducted by Whitaker (2010), more statistical sense?

As would be expected, the cognitive processing demands and the impact they have on test-retest performance are somewhat different for the Verbal Comprehension subtests. Although an overall progression effect was observed with these tasks, as shown in Table 14, subtests in these domains had some of the highest percentages of no change, with Vocabulary and Information having the highest percentages of no change (37% and 39%, respectively) of all subtests. Also, as shown in Table 5, the gains in these subtests were much smaller for a greater proportion of examinees than for the more novel subtests. Given what we know about the nature of the cognitive processes primarily involved in the performance of the Vocabulary and Information subtests (retrieval from long-term storage), these results are not surprising. Information and Vocabulary are measures of the ability to retrieve crystallized/acquired/learned knowledge from long-term storage. These tasks are not novel to the examinee and thus would not be expected within a neuropsychological framework to have large gains between time 1 and 2 because of the relatively small chance of an examinee making significant gains in crystallized/acquired knowledge in such a brief time. Minor variations likely result from increased or decreased efficiency in the retrieval of information from long-term storage (Whitaker, 2010).

Although many examinees had an improvement in time 2 scores, some negative score changes were apparent from time 1 to time 2 for all subtests. Within the neuropsychological framework, variations in performance, both increases and decreases, can be conceptualized primarily as variations in cognitive efficiency due to variations in the use of executive functions to direct effective performance. Specifically, instead of attributing these variations to random error of unexplainable origin, increases and decreases in performance could be conceptualized as the degree of efficiency an examinee has in the use of executive functions, as these are mental processes that cue the use of the other cognitive capacities to perform a task. For example, although an examinee is not likely to acquire or lose a significant amount of crystallized knowledge within the test-retest time frame, they may indeed demonstrate more or less efficiency in retrieving specific information from long-term storage. This is a subtle yet important difference, in that score differences from time 1 to time 2 are not a reflection of changes in word knowledge (vocabulary) as much as a reflection of changes in how efficiently an examinee is retrieving the knowledge of words from long-term storage. Retrieval, along with other executive functions, such as directing the monitoring of performance for accuracy, directing sustained attention to task, and directing set shifting are but a small sample of the various executive functions required for directing cognitive capacities to achieve successful performance.

Beyond providing a more neuropsychologically and statistically sound representation of variations in test performance from time 1 to time 2, the present model also provides clinicians with the ability to derive more meaningful information from their assessments to potentially drive the development of more meaningful interventions. Variations conceptualized as random error preclude subtest level analysis that could provide the highest level of specificity for intervention development. If clinicians interpret by psychometric strength alone, the strongest coefficient would be the full scale IQ. Interpreting only the full scale IQ, even when index scores are relatively consistent with one another, obscures potentially significant strengths and weaknesses that could be revealed via subtest level analysis/interpretation. Obscuring this information is ultimately a disservice to examinees and has the potential to lead to treatments that could at best be termed generic and at worst ineffectual. The results in Table 14 highlight these points. As Whitaker (2010) similarly found, there appears to be a lack of meaningful relationships between score change patterns based on the neuropsychological model and the reliability coefficient. Further, interpreting only the reliability coefficient obscures meaningful differences concerning the nature of score increases and decreases. For example, note the nearly identical correlation coefficients of .84 and .82 obtained for the Similarities and Picture Completion subtests, despite the markedly different patterns of score change exhibited for these two subtests (changes of 25%, 35%, and 45% for Similarities vs. 10%, 19%, and 71% for Picture Completion). Examining the percentage of change for these two subtests, it is difficult to accept the assertion of the psychometric model that performance on these two subtests is equal in stability over a 2- to 4-week period. Rather, the change data reveal that examinees were much more likely to have score increases and much less likely to have score decreases on Picture Completion than for Similarities.

Conversely, although the change patterns were nearly identical for the Digit Symbol-Coding and Object Assembly subtests (16%, 22%, and 62% for the Digit-Symbol Subtest vs. 18%, 20%, and 62% for the Object Assembly subtest), the correlation coefficients of .82 for Digit Symbol-Coding and .72 for Object Assembly suggest that Digit Symbol-Coding had a much more stable pattern of performance from time 1 to time 2 than Object Assembly.

As noted above, the current study builds upon the work of Whitaker (2010) in applying the neuropsychological performance consistency model first espoused by McCloskey (1990) to an adult population. It is interesting to interpret the results of the current study in light of what is known concerning cognitive theories of aging. In terms of the fluid/crystallized theory of cognitive aging (Horn and Cattell, 1966, 1967) one would expect to see comparatively fewer gains on fluid tasks than crystallized tasks in an older population (WAIS–III) than a younger population (WISC–IV). A sampling of the percentage of actual score increases on subtests generally believed to access fluid abilities did not conform to this hypothesized trend. The percentage of cases with score increases on the Block Design, Matrix Reasoning, and Picture Completion subtests was 59%, 48%, and 74%, respectively, for the WISC–IV test-retest sample, and 50%, 41%, and 71%, respectively, for the WAIS–III test-retest sample. Nonetheless, this is a somewhat difficult comparison to make methodologically because the populations were not stratified by age in the Whitaker study or the current study. Thus, for both test-retest samples, examinees at various ages, and thus at different stages of age-related cognitive development, were analyzed together. It can be conclusively stated, however, that there were significantly higher percentages of examinees with score increases on both the WISC–IV and WASI–III, and that in many cases, this change reflected a progression effect rather than regression effect.

Limitations and suggestions for future research.

The present study represents an extension of Whitaker's (2010) initial attempt to reconceptualize reliability and the manner in which intelligence test subtest reliability is presented. Future research is needed to refine the methodology and to extend the results. The present study borrows heavily from the research by Whitaker (2010); thus, many of the limitations continue to exist.

For example, the test-retest interval of 14 to 30 days is relatively short compared to test-retest intervals that may be present in nonexperimental settings. Thus, this may adversely impact the extent to which the results of this study can be generalized to the clinical setting. Future research could both increase the test-retest intervals and include multiple repeat administrations. It would be very interesting to see if third, fourth, and fifth administrations would have an overall progression effect noted in both the present study and that of Whitaker (2010). Further, if they did continue to result in the progression effect, will there be a point of diminishing returns in increased performance? Will there be variations, depending on the particular cognitive demands of the subtest? These are questions worthy of further research.

Future researchers could stratify samples based on age to examine whether test-retest reliability performance patterns are concordant with expected patterns, based on the neuropsychological literature. It also would be interesting to examine whether the data pattern of score progression also occurs in a preschool population, using the early childhood version of the Wechsler Scales (WPPSI–III). This might be particularly interesting, given the neuropsychological performance consistency model's conceptualization of score variation being driven the degree of efficiency in using executive functions, which are likely at their weakest stage of neural development during preschool years.

Further limitations and suggestions for future research relate to the samples used in these test-retest studies. According to the WAIS–III manual, exclusion criteria included, but were not limited to, memory problems, head injuries, and various medical and psychiatric conditions that may impact cognition. It would be interesting to examine the test-retest patterns in samples of individuals with impairments to determine whether they are concordant with current results. Due to the importance of executive functions within the neuropsychological framework, it may be especially beneficial to have studies conducted in populations with executive dysfunction, such as individuals diagnosed with

ADHD. Whitaker (2010) also called for studies to be conducted to determine test-retest performance for populations of learning disabled (LD) individuals, and specifically LD subtypes. Hain et al. (2009) developed a framework of learning disability subtypes based on cognitive, academic, and behavioral/social-emotional factors that might be especially appropriate as a basis for sample selection for such a study.

The most obvious extension would be to apply the methodology in this study to the Wechsler Assessment of Adult Intelligence – Fourth Edition (WAIS–IV, Wechsler, 2008). The development of the WAIS–IV represented significant changes in test structure and theoretical orientation in comparison to the WAIS–III (Lichtenberger, Sotelo-Dynega, & Kaufman, 2009). Examples of key revisions include updated norms, updated theoretical foundations, enhanced clinical utility, enhancement of fluid reasoning measurement and strengthening the overall framework of the test via factor analysis. Specifically, in terms of theoretical changes, fluid reasoning, working memory, and processing speed were given much greater attention than they received in the WISC–III. In addition, the WAIS–IV now has a more modern and conceptually clear scale structure, with the elimination of the older Verbal and Performance IQs and the adoption of the four-scale structure of the WISC–IV: Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed (Lichtenberger et al., 2009).

During the transition to the WAIS–IV, three new subtests were added in place of two that had been deleted from the WAIS–III. Specifically, Picture Arrangement and Object Assembly were deleted and the Visual Puzzles, Figure Weights, and Cancellation subtests were added. Visual Puzzles is part of the Perceptual Reasoning scale and is a timed subtest in which the examinee views a completed puzzle and the appropriate response option that reconstructs the puzzle. Figure Weights is also part of the Perceptual Reasoning scale and is a timed assessment in which the subject views a scale with missing weights and selects the response that keeps the scale balanced. Lastly, Cancellation is part of the Processing Speed Index and is a timed assessment in which the examinee scans and marks target shapes from a stimulus set (Lichtenberger et al., 2009).

As noted above, the results of this study were very consistent with results obtained from the performance consistency model Whitaker (2010) applied to the WISC–IV. When applying this same model and methodology to the WAIS–IV, it is predicted that results would again be similar. Regardless of whether the sample is grouped by total sample or in reference to subject performance to the mean, an overall progression effect rather than regression to the mean would be predicted. Specifically, regardless of scores in relation to the mean, most subtests would have a much greater frequency of positive than negative scores compared to what would be predicted, given the traditional psychometric model. Also, when grouping subjects in terms of their relation to the mean, a poor goodness of fit obtained via chi-square analysis would be expected between traditional psychometric reliability estimates and actual results obtained via the performance consistency model.

The most robust progression gains were generally noted on subtests within the Perceptual Reasoning and Processing Speed indices. These subtests require the most novel cognitive processing and, given the results of the current study, it would be predicted that the most robust gains would be noted in these subtests. As noted above, during initial exposure to these types of subtests, subjects are likely to develop problem-solving strategies for how to complete items. During subsequent exposure to the subtests, the subjects then have preexisting strategies to draw upon that they are able to refine and express more efficiently via executive functions that likely lead to performance gains. However, as noted above, subtests that are likely most susceptible to performance gains (Picture Arrangement and Object Assembly) have been deleted from the WAIS–IV. Although the newly adopted Figure Weights and Visual Puzzles would also be predicted to show large performance gains, they would likely not be quite as large as those evidenced in Picture Arrangement and Object Assembly on the WAIS–III simply because of the peculiarities of the task requirements and subtest design. During Picture Completion, the examinee is given 20 seconds to scan a picture and identify what essential aspect of the picture is missing. Progression gains on the second exposure to the task likely result from having already been exposed to the task, knowing what to look for, and noticing details that may have been missed on the initial scan. Although Figure Weights and Visual Puzzles are both novel tasks, they likely do not possess the same sensitivity to second exposure impact.

Similarly, the smallest effects in the present study were found on subtests that required more crystallized or codified knowledge. These subtests address more routinized information that is less likely to increase from time 1 to time 2 assessment. This same pattern of performance would be expected in WAIS–IV subtests, with subtests such as Vocabulary and Information, although still having an overall progression effect, having some of the lowest percentages of performance change.

In conclusion, the present study sought to extend the body of research in the field of neuropsychologically based performance models and their appropriateness in representing test-retest performance. Overall, this model was found to be more appropriate and useful in interpreting actual test results of examinees.

References

American Educational Research Association. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

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

Ballmaier, M., Kumar, M., & Elderkin-Thompson, V. (2003, June). *Cortical abnormalities in elderly depressed patients*. Lecture presented at the Human Brain Mapping Conference, New York, NY.

Baltes, P. B. (1997). On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as the foundation of developmental theory. *American Psychologist, 52*, 366-380.

Beauchamp, M. H., Dagher, A. Aston, J. A. D., & Doyon, J. (2003). Dynamic functional changes associated with cognitive skill learning of an adapted version of the Tower of London task. *NeuroImage 20*(3), 1649-1660.

Beebe, D. W., Pfiffner, L. J., & McBurnett, K. (2000). Evaluation of the validity of the WISC–III comprehension and picture arrangement subtests as measures of social intelligence. *Psychological Assessment, 12*, 97-101.

Berninger, V. (1992). A developmental neuropsychological perspective for reading and writing acquisition . *Educational Psychologist, 27*(4), 415-434.

Botwinick, J. (1977). Intellectual abilities. In J. E. Birren (Ed.), *Handbook of the psychology of aging* (pp. 580-605). New York, NY: Van Nostrand Reinhold.

Catron, D. W., & Thompson, C. C. (1979). Test-retest gains in WAIS scores after four retest intervals. *Journal of Clinical Psychology, 35*(2), 352-357.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(*3), 297-333.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137-163.

Daniel, M. H. (2007). Scatter and construct validity of FSIQ. *Applied Neuropsychology, 14*(4), 291-295.

Dumont, R. F., 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.

Elliot, C. D. (1990). *Differential Ability Scales (DAS): Introductory and technical handbook*. San Antonio, TX: Psychological Corporation.

Erickson, R. C. (1995). A review and critique of the process approach in neuropsychological assessment. *Neuropsychology Review*, *5*(4), 223-243.

Flanagan, D., & Kaufman, A. S. (2004). *Essentials of WISC–IV assessment*. New York, NY: Wiley.

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

Frederickson, N. (1999). The ACID test - or is it? *Educational Psychology in Practice, 15,* 2-8.

Goldberg, E. (2009). *The new executive brain*. New York, NY: Oxford.

Goldberg, E., & Costa, L. D. (1981). Hemispheric differences in the acquisition and use of descriptive systems. *Brain Language*, *14,* 144-173.

Gresham, F. M., & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future decisions. *School Psychology Quarterly, 12*, 249-267.

Grieve, S. M., Clark, C. R., & Gordon, E. (2003). The correlation between EEG, ERP and MRI in the same 110 normal human subjects [Abstract]. *Neuroimage, 19,* 32.

Hain, L. A. (2009). *Exploration of specific learning disability subtypes differentiated across cognitive, achievement, and emotional/behavioral variables.* (Unpublished doctoral dissertation). Philadelphia College of Osteopathic Medicine, Philadelphia, PA.

Hale, J. B., & Fiorello, C. (2001). Beyond the rhetoric of g: Intelligence testing guidelines for practitioners. *School Psychologist, 55*(4), 113-139.

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

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

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 39-683.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253-270.

Hubert, V., Beaunieux, H., Chételat, G., Platel, H., Landeau, B., Danion, J.-M., . . . Eustache, F. (2007). The dynamic network subserving the three phases of cognitive procedural learning. *Human Brain Mapping, 28*(12), 1415-1429.

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.

Kamiya, H., Umeda, K., Ozawa, S., & Manabe, T. (2002). Presynaptic Ca2+ entry is unchanged during hippocampal mossy fiber long-term potentiation. *Journal of Neuroscience 22*(24), 10524-10528.

Kaplan, E. (1998). A process approach to neuropsychological assessment. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function: Research, measurement, and practice* (pp. 127-167). Washington, DC: American Psychological Association.

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

Kaufman, A. S. (2003). Practice effects by guest author Alan Kaufman. Retrieved from http://www.speechandlanguage.com/cafe/13.asp

Kaufman, A. S., & Kaufman, N. L. (1983). *Manual for the Kaufman Assessment Battery for Children (K-ABC).* Circle Pines, MN: American Guidance Services.

Kaufman, A. S., & Kaufman, N. L. (1993). *Manual for the Kaufman Adolescent and Adult Intelligence Test (KAIT).* Circle Pines, MN: American Guidance Services.

Koziol, L. F., Budding, D. E., & Chidekel, D. (2010). Adaptation, expertise, and giftedness: Towards an understanding of cortical, subcortical, and cerebellar network contributions. *Cerebellum, 9*(4), 499-529.

Krenk, L., Rasmussen, L. S., Siersma, V. D., & Kehlet H. (2012). Short-term practice effects and variability in cognitive testing in a healthy elderly population. *Experimental Gerontology 47*(6), 432-436.

Leach, L., Kaplan, E., Rewilak, D., Richards, B., & Proulx, G.B. (2000). *Kaplan-Baycrest Neurocognitive Assessment (KBNA): Manual.* San Antonio, TX: Psychological Corporation.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). New York, NY: Oxford University Press.

Lichtenberger, E. O., Sotelo-Dynega, M., Kaufman, A. S. The Kaufman Assessment Battery for Children – Second edition. In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 61-93). Hoboken, NJ: John Wiley & Sons.

Luria, A. R. (1973). *The working brain*. Baltimore, MD: Penguin Books.

Majovski, L. V. (1997). Development of higher brain functions in children: Neural, cognitive, and behavioral perspectives. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 17-41). New York, NY: Plenum Press.

Mayes, S. D., & Calhoun, S. L. (2007). Wechsler Intelligence Scale for Children – Third and Fourth edition predictors of academic achievement in children with attention deficit hyperactivity disorder. *School Psychology Quarterly, 22*(2), 234-249.

Matarazzo, J. D., Carmody, T. P., & Jacobs, L. D. (1980). Test-retest reliability and stability of the WAIS: A literature review with implications for clinical practice. *Journal of Clinical Neuropsychology, 2*, 89-105.

McCarthy, D. A. (1972). *Manual for the McCarthy Scales of Children's Abilities.* New York, NY: Psychological Corporation.

McCloskey, G. (1990). Selecting and using early childhood rating scales. *Topics in Early Childhood Special Education, 10*(3), 39-65.

McCloskey, G. (2009). Clinical applications I: A neuropsychological approach to interpretation of the WAIS–IV and use of the WAIS–IV in learning disability assessments. In E. O. Lichtenberger & A. S. Kaufman (Eds.), *Essentials of WAIS–IV assessment* (pp. 208-244). Hoboken, NJ: Wiley.

McCloskey, G., & Maerlander, A. (2005). The WISC–IV Integrated. In A. Prifitera, D. Saklofske, & L. Weiss (Eds.), *WISC–IV clinical use and interpretation: Scientist practitioner perspectives* (pp. 101-149). San Diego, CA: Elsevier.

McCloskey, G., Perkins, L., & Van Divner, R. (2009). Assessment and intervention for executive function difficulties. New York, NY: Routledge Press.

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

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Miller, D. C. (2007). Essentials of school neuropsychological assessment. Hoboken, NJ: John Wiley & Sons.

Miller, D. C., & Hale, J. B. (2008). Neuropsychological applications of the WISC-IV and WISC–IV Integrated. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC–IV clinical assessment and intervention* (2nd ed., pp. 445-495). San Diego, CA: Elsevier.

Newman-Norlund, R. D., van Schie, H. T., van Zuijlen, A. M. J., & Bekkering, H. (2007). The human mirror neuron system more active during complementary compared with imitative action. *Nature Neuroscience, 10,* 817-818.

Poldrack, R. A., Sabb, F. W., Foerde, K. Tom, S. M., Asarnow, R. F., Bookheimer, S. Y.,& Knowlton, B. J. (2005). The neural correlates of motor skill automaticity. *Journal of Neuroscience 25*(22), 5356-5364.

Poldrack, R.A. & Willingham, D.B. (2006). Functional neuroimaging of skill learning. In R. Cabeza & A. Kingstone (Eds.), *Handbook of neuroimaging of cognition* (2nd ed., pp. 113-148). Cambridge, MA: MIT Press.

Posner, M. I., & Raichle, M. (1994). *Images of Mind.* New York, NY: Scientific American Library.

Prifitera, A., & Dersh, J. (1993). Base rates of WISC–III diagnostic subtest patterns among normal, learning disabled, and ADHD samples. In B. A. Bracken, R. S. McCallum (Eds.); *Wechsler Intelligence Scale for Children* (3rd ed., pp. 43-55). Brandon, VT: Clinical Psychology Publishing Co; 43-55.

Psychological Corporation. (1997). *WAIS–III integrated technical and interpretative manual*. San Antonio, TX: Author.

Psychological Corporation. (2004). *WISC–IV integrated technical and interpretative manual*. San Antonio, TX: Author.

Raichle, M. E., Fiez, J. A., Videen, T. O., Macleod, J. V., Pardo, P., Fox, T., & Petersen, S. E. (1994). Practice-related changes in human brain functional anatomy during nonmotor learning. *Cerebral Cortex, 4,* 26.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Rettmann, M. E., Prince, J. L., & Resnick, S. M. (2003, June). *Analysis of sulcal shape changes associated with aging*. Lecture presented at the Human Brain Mapping Conference, New York, NY.

Roebroeck, M. E., Harlaar, J., Lankhorst, G. J., Hayes, K. W., Matyas, T. A., Keating, J. L., & Greenwood, K. M. (1993). The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Physical Therapy 73*(6), 386-401.

Roland, P. E. (1993). Brain activation. New York, NY: John Wiley and Sons.

Ryan, J. J., Glass, L. A., & Bartels, J. M. (2010). Stability of the WISC-IV in a sample of elementary and middle school children. *Applied Neuropsychology 17*(1), 68-72.

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment in special and inclusive education* (9th ed.). Boston, MA: Houghton Mifflin.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications*. La Mesa, CA: Author.

Shatz, M. W. (1981). WAIS practice effects in clinical neuropsychology. *Journal of Clinical Neuropsychology, 3,* 171-191.

Staines, W. R., Padilla, M., & Knight, R. T. (2002). Frontal-parietal event-related potential changes associated with practicing a novel visuomotor task. *Cognitive Brain Research*, *13,* 195-202.

Subkoviak, M. J. (1980). Decision-consistency approaches. In R. E. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 129-185). Baltimore, MD: Johns Hopkins University Press.

Suen, H. K., & Lie, P. (2007). Classical versus G-theory of measurement. *Educational Measurement, 4,* 3-2.

Taki, Y., Groto, R., Evans, A., Zidenbos, A., Neelin, P., Lerch, J., . . . Ono, S. (2003, June). *Voxel based morphology of age related structural change of grey matter for each decade in normal male subjects* . Lecture presented at the Human Brain Mapping Conference, New York, NY.

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

Thorvaldsson, V., Hofer, S. M., Berg, S., & Johansson, B. (2006). Effects of repeated testing in a longitudinal age-homogeneous study of cognitive aging. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences 61*(6), P348-P354.

Tuma, J. M., & Appelbaum, A. S. (1980). Reliability and practice effects of WISC-R IQ estimates in a normal population. *Educational and Psychological Measurement 40*(3), 671-678.

Van Der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement, 4(4),* 469-492.

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

Watkins, M., & 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.

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., Glutting, J. J., & Lei, P. W. (2007). Construct validity of the WISC–III for a national sample of Native American students. *Applied Neuropsychology, 14,* 13-20.

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

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997b). 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., & Smith, L. G. (2013). Long-term stability of the wechsler Intelligence Scale for Children – Fourth Edition. *Psychological Assessment 25*(2), 477-483.

Watkins, & 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, D. (1955). *WAIS manual*. New York, NY: Psychological Corporation.

Wechsler, D. (1967). *Wechsler Preschool and Primary Scale of Intelligence*. New York, NY: Psychological Corporation.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children – Revised (WISC–R).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale – Revised (WAIS–R).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence – Revised (WPPSI–R).* San Antonio, TX: Psychological Corporation.

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

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale – Third Edition (WAIS–III)*. San Antonio, TX: Harcourt Assessment.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence – Third Edition (WPPSI–III).* San Antonio, TX: Harcourt Assessment.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children –Fourth Edition (WISC–IV).* San Antonio, TX: Harcourt Assessment.

Whitaker, J. (2010). *Using a performance consistency model to explain variations in test-retest performance* (Unpublished doctoral dissertation). Philadelphia College of Osteopathic Medicine, Philadelphia, PA.