2011

Using a Performance Consistency Model to Explain Variations in Test-Retest Performance

James S. Whitaker
Philadelphia College of Osteopathic Medicine, james.whitaker@pearson.com

Recommended Citation

Whitaker, James S., "Using a Performance Consistency Model to Explain Variations in Test-Retest Performance" (2011). *PCOM Psychology Dissertations.* Paper 178.

Philadelphia College of Osteopathic Medicine

Department of Psychology

USING A PERFORMANCE CONSISTENCY MODEL TO EXPLAIN VARIATIONS IN TEST-RETEST PERFORMANCE

James S. Whitaker

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Psychology

April 2011

PHILADELPHIA COLLEGE OF OSTEOPATHIC MEDICINE
DEPARTMENT OF PSYCHOLOGY

Dissertation Approval

This is to certify that the thesis presented to us by James Whitaker

on the 25th day of May, 2010, 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.

Committee Members' Signatures:

George McCloskey, Ph.D., Chairperson

James Brad Hale, Ph.D.

Dr. Candice Stefanou

Robert A. DiTomasso, Ph.D., ABPP, Chair, Department of Psychology

Acknowledgments

I wish to thank George McCloskey for choosing me to bring to life his ideas on test reliability and brain-behavior relationships. I could never have completed this project without George’s endless support and guidance. Thank you to Candice Stefanou for always believing in me and for planting the seeds that led me to PCOM. A very special thanks also to Brad Hale for his support and expert guidance during this project. Thank you to Bruce Kelly for teaching me how to write and for making me a good clinician. To my mother, who sacrificed so much so that I could pursue my goals, may this dissertation always serve as a reminder of what you have accomplished. And to my father, thanks for teaching me to love learning about new things. To Olivia, my precious little girl, no accomplishment is greater than all that I received the day you were born; thank you for coming into my life. And to my wife, Mindy, I know that this project was made so much more difficult as a result of the tragedy and heartache we experienced during its completion. No words can express my gratitude for your support and constant demonstration of strength and courage. I dedicate this work in loving memory of our sister, Amy Williams, who lived and died with courage.

Abstract

Subtest-level interpretation of intelligence tests is necessary for understanding the relationship between cognitive deficiencies and academic problems and for designing interventions based on assessment results. However, the practice of subtest interpretation continues to be discouraged by those who claim that subtests have poor reliability and thus minimal interpretative power. This perception of subtest instability may be the result of misguided conceptions of reliability and not actual properties of subtests. With this in mind, the present study sought to determine the extent to which a neuropsychologically based performance model fit WISC-IV subtest test-retest data and offered an alternate means of understanding and interpreting the concept of subtest reliability. Higher rates of score progression versus regression were demonstrated for all subtests regardless of whether or not time 1 scores were above or below the mean. Rates of score increases from time 1 to time 2 varied based on the neuropsychological basis of the task being assessed.   Results suggest that a neuropsychologically based performance model is superior to a traditional psychometric model for representing WISC-IV subtest reliability and the manner in which individuals use their brains when they engage tasks.

Table of Contents

List of Tables

Chapter 1

Introduction

The history of Wechsler Scales interpretation has been marked by ongoing debate between Full Scale IQ (FSIQ) advocates and those who favor analysis of individual strengths and weaknesses at the index, subtest, and item levels. Recognizing the tendency of global (FSIQ) and index scores to obscure clinically relevant information, advocates of subtest-level interpretation argue that this approach is necessary for understanding children's cognitive deficits and providing interventions. However, subtest-level interpretation continues to be criticized on the basis that subtests have poor reliability and limited interpretative power. These criticisms may be the result of misguided conceptions of reliability that fail to consider the manner in which individuals use their brains when they engage tasks. With this in mind, the present study was designed to accomplish two goals. The first was to determine the extent to which a neuropsychologically based performance model could account for actual subtest test-retest findings for one of the Wechsler Scales, the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003). The second was to determine the effectiveness of a neuropsychologically based performance model compared to traditional psychometric procedures in terms of the type of information it provides test consumers regarding WISC-IV subtest reliability.

Literature Review

Review of the History and Methods of Wechsler Scales Interpretation

Global composite intelligence test score interpretation, referred to in this study as FSIQ, has its roots in Spearman's (1904) factor analytic studies that suggested a single factor common to all the cognitive tasks he studied. Spearman labeled this underlying factor as *g*. Over time, *g* has come to be viewed as the quintessential measure of overall intellectual ability. Currently, *g* is thought to be represented and quantified via the FSIQ in tests such as the WISC-IV, the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008), and the Wechsler Preschool and Primary Scale of Intelligence-Third Edition (WPPSI; Wechsler, 2002). *G* is conceptualized as a trait that is stable over time and not easily modified by educational or environmental interventions (Gottfredson, 1997).

FSIQ proponents claim that scores below the global level have less interpretative power primarily because of poor reliability (McDermott, Fantuzzo & Glutting, 1990; McDermott, Fantuzzo, Glutting, Watkins & Baggaley, 1992). They also cite the stability of global scores in cases where index or subtest scores fluctuate from one test administration to the next and argue that FSIQ, as a representation of *g*, is the only valid and reliable means for characterizing cognitive ability. As evidence for this latter claim, they cite studies documenting the relationship between *g* and overall life success, including educational and occupational attainment, marital satisfaction, and emotional health (Buckholdt, 2001; Gottfredson, 1997; Kranzler, 2001).

From this perspective, variation in task performance at the subtest level represents the effect of sources of error and the individual task's inability to accurately reflect the *g* that is thought to be accurately reflected in the global FSIQ. In Gottfredson's (1998) words:

> Because every mental test is "contaminated" by the effects of specific mental skills, no single test measures only *g*. . . . The scores from IQ tests . . . contain some "impurities." . . . For most purposes, these "impurities" make no practical difference, and g and IQ can be used interchangeably. (p. 26)

As thus intimated by Gottfredson (1998), the interpretation of FSIQ requires adherence to the assumption that individual performance across the multiple subtests and indexes from which it is derived is relatively uniform. In clinical practice, we know that this is not always the case, and some have argued against the interpretation of FSIQ in cases of extreme index and subtest score variability (e.g., Fiorello, Hale, McGrath, Ryan & Quinn, 2001; Kaufman & Flanagan, 2009; Prifitera, Weiss, & Saklofske, 1998).

Proponents of a multiple factors approach to interpretation utilize index scores that are derived from combinations of subtests thought to measure the same cognitive capacities (Kaufman & Flanagan, 2009; Lichtenberger & Kaufman, 2009). Unlike Gottfredson and like-minded theorists, these proponents of index and subtest-level interpretation view variations in task performance as potentially clinically meaningful rather than merely "contamination" of *g*. Those who favor this position suggest that the index and subtest score variability demonstrated by many children reduces both the descriptive and predictive validity of FSIQ (Fiorello et al., 2007).

Like FSIQ interpretation, index or factor-level interpretation also relies on the assumption that individuals utilize in similar manner closely related groups of cognitive functions when completing index-specific tasks. However, this proposition is not supported by the literature, which suggests that intelligence test indexes measure multiple cognitive functions, even within a specific index (Flanagan & Kaufman, 2009; Hale & Fiorello, 2004; McCloskey, 2009). Recognizing this literature base, some psychologists advocate for subtest interpretation in place of or as a supplement to index and FSIQ interpretation (Weiss, Saklofske, Prifitera & Holdnak, 2006).

Several models for subtest-level interpretation exist. Sattler (2001) provided a procedure similar to that initially proposed by Kaufman (1979) for identifying cognitive strengths and weaknesses based on deviations from the arithmetic mean of a group of subtests. Flanagan and Kaufman (2009) offered a model derived from Kaufman's (1979; 1994) "intelligent testing" approach. The model applies both nomothetic and idiographic procedures in the interpretation of global, index, and subtest scores. While they do not promote individual subtest interpretation on the grounds that single subtests are not psychometrically sound, Flanagan and Kauffman do recommend the interpretation of subtest clusters based on their shared abilities identified in the Catttel-Horn-Carroll (CHC) theory of the structure of cognitive abilities.

Like those who subscribe to subtest profile interpretation, advocates of the process approach to psychological assessment also propose that index and FSIQ interpretation can mask clinically meaningful information and lead to inaccurate characterizations of cognitive ability (Kaplan, 1988; Kaplan, Fein, Morris, Kramer &

Delis, 1999; McCloskey, 2009a; McCloskey & Maerlender, 2005). With foundations in a Lurian neuropsychological theory (Luria, 1973), the process approach is based on the work of Kaplan and colleagues (Kaplan, 1988; Kaplan et al., 1999) and involves interpretation beyond the level of subtest performance (i.e., average, below average, etc.). Subtest scores are conceptualized as products of multiple sources of influence, including the format of the tasks, the unique cognitive processes and abilities utilized by the child, and the specific strategies used by the child to perform the task (McCloskey & Maerlender, 2005).

McCloskey and Maerlender (2005) have identified the following interconnected principles which serve as the basis for the process approach and for understanding the type of information cognitive assessment yields: (1) intelligence subtests are multifactorial tasks that involve a complex interaction of many neuropsychological processes; (2) identifying the cognitive processes that contribute to successful or unsuccessful task performance allows for the identification of the source of cognitive deficiencies and strengths and the establishment of brain-behavior relationships; (3) the cognitive skills utilized during task completion may vary from one individual to another based on how that individual responds to the input, internal processing, and output demands of the task; (4) careful and systematic observation of performance during completion of a subtest is necessary for understanding how the individual achieved his/her score; and (5) observations during task completion, including analyses of error patterns, can lead to confirmation or refutation of hypotheses regarding the origin of deficits.

Several models of intelligence-test interpretation represent extensions of Kaplan's (1988) process approach. Hale and Fiorello (2004) suggested an interpretative cycle that begins with analysis of global levels of performance and includes subtest interpretation when significant variability renders FSIQ invalid (Fiorello et al., 2007). The focus of their idiographic assessment is the cognitive processes necessary for task completion. Recognizing the multifactorial nature of subtests, they recommend conducting *demands analyses* for subtests representing cognitive strengths and weaknesses. The demands analysis includes an assessment of the input, processing, and output demands of each subtest for the individual child.

McCloskey (2009) also provided a neuropsychologically oriented interpretative levels model focusing on the interpretation of clinical clusters, subtests, items, and the cognitive capacities required to complete tasks. Like Hale and Fiorello (2004), he stressed the importance of careful observation during assessment to facilitate an understanding of the unique cognitive processes utilized by individuals during the completion of subtest tasks.

Cognitive Capacities Measured by the WISC-IV Subtests

The WISC-IV (Wechsler, 2003) is designed for use with children ages 6 years, 0 months through 16 years, 11 months. The instrument is composed of 15 subtests, 10 core and 5 supplemental. The 10 core subtests yield 4 indexes: the Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI). The FSIQ composite is derived from the 10 core subtests.

Also provided for select subtests are process scores, which allow further examination of the primary and secondary cognitive processes involved in task completion.

The VCI consists of three core subtests (Vocabulary, Similarities, and Comprehension), and two supplemental subtests (Word Reasoning and Information). The PRI also consists of three core subtests (Block Design, Picture Concepts, and Matrix Reasoning), and one supplemental subtest (Picture Completion). The WMI includes two core subtests (Digit Span and Letter-Number Sequencing), and one supplemental (Arithmetic). The PSI is composed of two core subtests (Coding and Symbol Search), and one supplemental subtest (Cancellation). Table 1 provides a brief description of the WISC-IV core and supplemental subtests.

Table 1

Descriptions of WISC-IV Core and Supplemental Subtests

Subtest	Description
Vocabulary	The examinee names pictures and provides verbal definitions of words.
Similarities	The examinee verbally describes the concrete and abstract similarities between sets of concepts or objects.
Comprehension	The examinee provides oral responses to questions requiring common sense and knowledge of conventional standards of behavior.
Information	The examinee provides oral responses to questions assessing general knowledge.
Word Reasoning	The examinee identifies a target word based on a series of clues.
Block Design	The examinee constructs three-dimensional block designs using a model within the allotted time.
Matrix Reasoning	The examinee performs a matrix analogy task requiring determination of part-whole relationships.
Picture Concepts	The examinee identifies the common characteristic among two or three rows of pictures.
Picture Completion	The examinee views a series of pictures and identifies the essential part they are missing within the allotted time.
Digit Span	The examinee repeats a series of orally presented digits both backwards and forward.
Letter-Number Sequencing	The examinee is required to listen to a series of numbers and letters and then repeat them in ascending and alphabetical order.
Arithmetic	The examinee mentally solves a series of orally presented arithmetic problems within the allotted time.
Coding	The examinee uses a grid to copy geometrical symbols within a specified time limit.
Symbol Search	The examinee scans an array of symbols to determine the presence or absence of a target symbol within a specified time limit.
Cancellation	The examinee scans an array of pictures presented in random and nonrandom fashion and identifies target pictures within the allotted time.

Note. Subtests in italics are supplemental.

Several authors (McCloskey, 2009; Miller, 2007; Miller & Hale, 2008) have identified both primary and secondary cognitive and neuropsychological processes likely assessed by the WISC-IV subtests. The primary cognitive processes are those constructs that the subtest is designed to measure. Secondary processes are the cognitive constructs that may not be the focus of the assessment task but that may support successful task execution. Poor subtest performance can result from lack of effective use of primary or secondary capacities or from a combination of both. As provided by McCloskey (2009) and Miller and Hale (2008), Table 2 provides a brief overview of the hypothesized primary and secondary cognitive capacities assessed by the WISC-IV subtests.

Table 2

Cognitive Capacities Assessed by the WISC-IV Subtests

	WISC-IV Core and Supplemental Subtests														
Cognitive Capacity	VC	SI	Co	In	WR	BD	MR	Pcn	Pcm	Ds	Ls	Ar	Cd	SS	Ca
Executive Functions	S	S	S		S	S	P	P	S	S	S	S	S	S	S
Memory Functions	P	S		P	P				S	P	P	P			
Auditory Perception	S	S	S	S						S	S	S			
Language Functions	P		P						S						
Reasoning Ability		P	P		S	P	P	P	P						
Visuomotor Processing Speed						S							P	S	
Visual Perception						P	S	P	S				S	S	S
Visual Processing Speed														P	P

Note. P = Primary; S = Secondary

Executive functions refer to the broad category of cognitive processes responsible for cueing and directing mental activities and behaviors, such as attention, effort, problem-solving, and response inhibition. Executive functions appear to be a key mediator in the successful performance of all WISC-IV subtests. Memory functions refer to working memory as well as to the initial encoding of and retrieval of information from long-term storage.

Auditory perception refers to the broad array of cognitive capacities involved in the accurate perception and discrimination of speech sounds, the comprehension of grammar and syntax, and the efficient registration of auditory information. Language functions refer to both expressive and receptive language capabilities. The term *reasoning*, in the context of subtest performance, refers to the examinee's ability to think abstractly about verbally mediated information and nonmeaningful, visual stimuli.

Visual perception refers to the broad category of visual processes involved in the accurate representation of visual stimuli and the ability to detect similarities and differences in visual stimuli. Visual processing speed refers to the examinee's ability to efficiently integrate visual and motor processes when completing tasks such as Block Design.

Rationale for and Criticisms of Subtest-Level Interpretation

Proponents of subtest-level analysis point out not only that global scale interpretation obscures clinically meaningful information but also that this approach has little diagnostic or treatment validity (Lezak, 1988). In contrast, numerous studies have demonstrated the link between cognitive processes and academic skills (Flanagan, Ortiz, Alfonso & Mascolo, 2002; Floyd, Evans & McGrew, 2003) and various forms of psychopathology (Hain, Hale, & Kendorski, 2009; Hale & Fiorello, 2004). Furthermore, it is apparent that examination of cognitive strengths and weaknesses, which can be accomplished only via subtest and item-level analysis, is necessary for understanding a child's learning problems and for developing interventions (Fiorello et al., 2001; Hale & Fiorello, 2004; Hale, Fiorello, Kavanagh, Hoeppner, & Gaither, 2001).

Nevertheless, subtest-level interpretation continues to be discouraged by those who adhere to psychometric traditions on the basis that subtests are less reliable than index or global scores (McDermott et al., 1990; McDermott et al., 1992) and more susceptible to measurement error (Macmann & Barnett, 1997). In contrast to this view, McCloskey (2009a) pointed out that the perception of subtest instability may be the result of misguided conceptions of reliability and not of actual properties of subtests.

Despite evidence of the multifactorial nature of intelligence tests, traditional psychometric methods for estimating intelligence-test reliability continue to rely on the questionable assumptions that Wechsler subtests measure specific, stable cognitive traits and that re-administrations of a subtest should produce the same results if the test is reliable. In the traditional psychometric model, variations in test-retest performance that may be related to factors specific to the internal mental states of the examinee and/or the application of the examinee's mental capacities with the specific format of the test materials are attributed to measurement error. Any variations from a static level of performance on first and subsequent administrations of the same task therefore are viewed as measurement error. Literal interpretation of these sources of variation as "measurement error" that produces undesirable and/or uninterpretable consequences is pointed to as evidence that the task is an unreliable source of information about the examinee's cognitive capacities.

Neuropsychology of Practice-Related Changes in Intelligence-Test Performance

The proposition that variable levels of change in performance from time 1 to time 2 are the result of undesirable measurement error is not consistent with the neuropsychological literature base that has examined performance on repeated administrations of a task. Numerous studies suggest that increases in cognitive efficiency, or what is commonly referred to as the practice effect, contribute to increases in performance on repeated measures of cognitive functioning (Catron & Thompson, 1979; Kaufman, 2003; Matarazzo, Carmody, & Jacobs, 1980). From a neuropsychological perspective, these increases in performance can be explained as resulting from changes in brain functioning that promote the learning of novel tasks and more efficient execution of previously learned skills. This proposition is consistent with neuroimaging studies indicating differences in cerebral activation patterns based on both task demands and previous exposure to the task (Bever & Chiarello, 1974; Gold, Berman, Randolf, Goldberg & Weinberger, 1996; Henson, Shallice & Dolan, 2000; Martin, Wiggs & Weisberg, 1997).

Clearly articulated in the seminal work of Goldberg and Costa (1981), the novelty-routinization hypothesis of hemispheric specialization suggests that the right hemisphere is more actively involved in the processing of ambiguous or novel information while the left hemisphere specializes in the processing of automatic, familiar information for which specific mental representations exist. A key phrase in the preceding sentence is “more actively involved,” as Goldberg and Costa were not implying task-specific localization but rather that both hemispheres are interconnected, with the psychological process involved determining the degree of hemisphere involvement (Hale & Fiorello, 2004). In addition, Goldberg (2001) notes that executive processes are more relevant for right hemisphere functions than for left hemisphere ones.

Goldberg and Costa's (1981) novelty-routinization hypothesis is based on the earlier work of Luria (1973), who suggested hemispheric specialization based on the internal organization and representation of information (Majovski, 1997). Luria also postulated that cognitive skill acquisition involves a gradual shift from anterior to posterior regions of the brain. Luria described three principal functional units in the brain. The first functional unit includes the reticular system and related structures and is responsible for the maintenance of tone or waking (Hale & Fiorello, 2004). The second functional unit is devoted to receiving, storing, and analyzing information and is housed in the posterior occipital, parietal, and temporal regions of the brain. Luria's third functional unit is the frontal lobes, which are responsible for the regulation of almost all aspects of mental activity (Hale & Fiorello, 2004).

Among Luria's principal functional units, the initial stages of learning a new task are characterized by greater use of anterior brain regions, specifically the frontal lobes. However, once a skill is mastered, posterior regions of the brain become more important in performing the learned task (Goldberg, 2001; Goldberg, Harner, Lovell, Podell, & Riggio, 1994; Hale & Fiorello, 2004).

Using neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), numerous researchers have attempted to correlate cerebral blood-flow patterns with right-left, anterior-posterior transitions during novel and familiar task completion and under conditions of repeated task exposure. Cerebral blood-flow levels are assumed to correlate with neural activity (Goldberg, 2009).

In a study designed to investigate practice-induced improvements in performance, Raichle et al. (1994) used PET to study activation patterns during the naïve and practiced performance of a verbal-response selection task. In the naïve condition of the study, participants were asked to say an appropriate verb for each visually presented noun from a list of 40 nouns. In another condition of the study, called the practice condition, a different group of subjects were given the list of nouns, asked to identify a verb to go with each noun on the list, and given 15 minutes to rehearse the noun-verb association list they generated. In the novel condition, subjects experienced the naïve condition first and then were given a new list of 40 nouns and asked to generate a verb for each, as in the naïve condition. Results indicated that the pattern of activation present during the naïve condition, the anterior cingulate, the left prefrontal and left posterior temporal cortices, and the right cerebellar cortices, all but disappeared during the practice condition and was partially reactivated during the novel condition. Furthermore, in contrast to the naïve and novel conditions, the practice condition was associated with significantly greater activation in the left medial occipital region.

Martin, Wiggs, and Weisberg (1997) studied the regional cerebral blood-flow patterns associated with learning two different sets of meaningful words, nonsense words, real objects, and nonsense objects. Performance during the first presentation of items was associated with activation of the right mesiotemporal structures, but this activation decreased during the second presentation. Activation of the left mesiotemporal structures was evident during both learning trials. According to Goldberg (2009), the finding of decreased right hemisphere activation during the second presentation of novel items suggests that cerebral activation patterns are related to general characteristics of learning and not to the learning of specific items. In other words, even though the Trial 2 items were different, the nature of the task was the same and thus no longer novel to the participants.

The findings of Martin, Wiggs, and Weisberg (1997) and Raichle et al. (1994) also provide compelling information regarding the localization of language functions in the brain. Contrary to traditional conceptions of language as the primary responsibility of the left hemisphere, both studies indicate that linguistic information is also processed in the right hemisphere, provided that the task requiring linguistic processing is novel.

In studies examining cerebral blood flow patterns associated with facial and symbol recognition, novelty was associated with right but not left activation of the hippocampal and parahippocampal structures (Tulving, Markowitsch, Craik, Hiabib & Houle, 1996) and right occipital regions (Henson et al., 2000). In contrast, a study examining perceptual decision making with easily recognizable items indicated greater activation of the left dorsolateral prefrontal regions (Heekeren, Marrett, Bandettini & Ungerleider, 2004).

In a study designed to examine the right-left and anterior-posterior transition, Gold et al. (1996) examined cerebral blood-flow patterns associated with learning a task requiring delayed response and alternation. Not surprising given the executive demands of this task, results indicated frontal lobe activation during both the early and later stages of task learning. However, this activation was significantly greater during early- as compared to late-stage learning. Also evident was a shift from right to left frontal lobe activation as participants became more familiar with the task.

The findings of Gold et al. (1996) are similar to those of Shadmehr and Holcomb (1997), who demonstrated greater right prefrontal activation during early but not later stages of learning a complex motor skill. Later stage learning was associated with more significant activation of the left posterior parietal cortex. Similarly, the results of Staines, Padilla, and Knight (2002) suggest both a right to left and an anterior to posterior transition during the learning of a visuomotor task.

Variability in brain function can account in a meaningful way also for decreases in performance when retested with the same task. As suggested earlier, executive functions mediate important performance variables, including sustained attention, effort, and motivation. Minor variations in retest performance can result from variation in the use of these executive-function capacities (Barkley, 2006; Denkla, 2007; McCloskey, 2009a; McCloskey, 2009b; McCloskey, Perkins, & VanDiviner, 2009).

Ineffective performance at time 1 can reduce a score, and this lowering effect can be overcome by more efficient use of executive functions on the second testing. Conversely, performance decreases from time 1 to time 2 can result from inefficient engagement of executive functions. While the effect of variations in cognitive efficiency would produce effects similar those proposed in a traditional psychometric model of reliability, the source of variability is attributed to neuropsychologically based brain states rather than to a confluence of random error factors.

Studies Addressing Practice-Related Changes in Intelligence Test-Retest Performance

Several variables, including the test-retest interval, task novelty, motor speed requirements, and examinee age (Kaufman, 2003; Lezak, Howieson & Loring, 2004; Shatz, 1981), have been examined as potential mediators of the degree of practice effects observed on intelligence tests. To assess the differential influence of practice effects at different test-retest intervals, Catron and Thompson (1979) administered the WAIS (Wechsler, 1955) on two occasions to four groups of college students at 1-week, 1-month, 2-month, and 4-month intervals. Results indicated that the largest increases in Verbal (VIQ), Performance (PIQ), and FSIQ occurred at 1 week and the smallest at 4 months. The average standard score increases in VIQ at 1 week, 1 month, 2 months, and 4 months were as follows: 4.7, 1.8, 2.3, and .8. The average standard score increases in PIQ and FSIQ at 1 week, 1 month, 2 months, and 4 months were as follows: 11.4, 9.8, 8.7, and 8.0 and 8.0, 5.7, 5.4, and 4.2, respectively.

To investigate the differential impact of practice effects on Wechsler's performance and verbal scales, Kaufman (2003) analyzed the 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). Results indicated higher test-retest standard score gains in PIQ and FSIQ versus VIQ. With mean age intervals ranging from 5 to 50 years and mean test-retest intervals of 3 to 11 weeks, the median gain on VIQ for these instruments was 3.2, while the median gains for PIQ and FSIQ were 9.0 and 6.8, respectively.

A pattern of differential practice effects similar to those observed on the Wechsler scales has also been demonstrated on other intelligence tests. The average gain of approximately 7 points on the Wechsler FSIQ noted by Kaufman (2003) also has been observed for the global index scores for 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).

The average gain on the Simultaneous Processing scale of the KABC (Kaufman & Kaufman, 1983), which resembles Wechsler's PIQ, is 6.5 compared to 2.5 on the Achievement scale, which is similar to the VIQ (Kaufman, 2003). On the SB-IV (Thorndike, Hagen & Sattler, 1986), gains on the Abstract/Visual Reasoning scale averaged 7.5 to 8 points, while gains on the Verbal Reasoning scale averaged 5 points. Finally, on the KAIT (Kaufman & Kaufman, 1993), practice effects produced an average 7-point gain in Fluid IQ compared to a 4.5 gain in Crystallized IQ, which is derived from subtests similar to those that make up Wechsler's VIQ.

Traditional Conceptions of Reliability

In contrast to performance changes based on brain-function adaptation to tasks, the traditional psychometric conception of test reliability can be broadly defined as the consistency of the measure or the extent to which the measure yields the same results on repeated trials (Anastasi & Urbina, 1997). Nunnally and Bernstein (1994) offer a more complete definition of reliability as the extent to which measurements can be replicated with different examiners, with alternative instruments designed to measure the same thing, and under circumstances where minor variations exist in the conditions of measurement. The current methods for conceptualizing, calculating, and presenting intelligence-test reliability data are based on test theory concepts, such as regression to the mean, error variance, and true scores.

Traditional psychometric theory is used when estimating the reliability of testing instruments. Traditional psychometric theory posits that scores on testing instruments are composed of two elements: true score variance and error variance. A true score is a hypothetical concept and, as such, can never be obtained or directly measured, but rather can be approximated through multiple administrations of the same test. The true score, therefore, can be conceptualized as the mean of the distribution of scores one would achieve if tested repeatedly with the same instrument (Nunnally & Bernstein, 1994). The true score can also be viewed as one’s true ability level or actual level of the characteristic of interest (e.g., true level of intelligence). This true score is based not only on measurement error, but also in the mathematical fact that as obtained scores deviate more from the mean, subsequent scores will show regression to the mean (Anastasi & Urbina, 1997).

Anastasi and Urbina (1997) broadly define error variance as any influencing factor that is not relevant to the purpose of the testing instrument. Nunnally and Bernstein (1994) identified content sampling as the primary source of measurement error within a test and also suggested that variance can result from multiple random factors, including examinee fatigue and administration errors.

Within the context of traditional psychometric theory, error, regardless of its source, is important to consider and quantify because as the amount of measurement error increases, the reliability of the instrument decreases.

An important underlying assumption of traditional psychometric theory is that all variance in an individual’s performance in multiple administrations of the same test is the result of measurement error and not variation in the mental capacities that are being assessed. This assumption is critical for consideration in that it limits greatly the utility of the psychometric theory conception of reliability when attempting to describe human performance patterns on intelligence tests.

Methods for Estimating Reliability

There are three conceptual approaches to determining the reliability of an instrument that are discussed in the traditional psychometric literature and in test manuals: internal consistency, test-retest, and interrater reliability.

Internal consistency.

Internal consistency methods provide an estimation of the consistency of scores across test items assumed to measure the same construct. Internal consistency reliability provides an indication of both test stability and the precision with which the construct of interest has been measured (Anastasi & Urbina, 1997). Methods to determine internal consistency reliability include alternate forms and the split-half method, for which a third method of internal consistency has been derived called coefficient/Cronbach's alpha.

In the alternate forms approach, two instruments measuring the same attribute and containing nonoverlapping sets of items are administered to the same individuals on two occasions, and their scores are correlated (Anastasi & Urbina, 1997; McDonald, 1999). To be considered alternate and equivalent, the two test forms must measure the same trait to the same extent and must be standardized on the same population.

The second method for determining internal consistency, referred to as the split-half method, involves creating two equivalent forms of a test measuring the same skill by dividing the test in half. Both tests are then administered, and the scores are correlated as in the alternate forms approach. If the test design is such that the easier items are presented first and the more difficult items last, the use of procedures to ensure that both forms of the test contain easy and difficult items will be necessary.
Coefficient/Cronbach's alpha is a variant of the split-half method. It is derived by averaging all possible split-half correlations within the items that comprise a subtest (Cronbach, 1951).

Test-retest.

Test-retest reliability is generally acknowledged as a measure of test score consistency over a short period of time, such as 2 weeks, while stability refers to the consistency of scores over long periods of time (Anastasi & Urbina, 1997). However, the test manuals for most widely used intelligence tests, including the Wechsler Scales (Wechsler, 1997, 2002, 2003), refer to estimates of reliability based on short test-retest intervals as stability coefficients. To calculate stability coefficients, test developers administer the test to a subset of individuals from the standardization sample on two occasions and then correlate the two sets of test scores.

Interrater.

Interscorer or interrater reliability refers to the extent to which different assessors provide similar scores or ratings when using the same instrument or when observing the same person (Anastasi & Urbina, 1997). There are primarily two methods for determining interrater reliability: the correlational method and percentage of agreement.

The correlational method is similar to that employed when determining test-retest reliability. Two individuals administer the same instrument, and the scores they assign are then correlated.

Generally speaking, the percentage-of-agreement approach involves calculating the proportion of agreement between ratings by the same person on the same instrument at different times or by different persons on the same instrument.

Current Procedures for Estimating Intelligence-Test Stability and Reliability

Because an individual's true score is not known, reliability cannot be directly quantified. To address this problem, traditional psychometric theory proposes mathematical constructs to quantitatively estimate and represent the amount of error associated with observed scores. The three most common of these mathematical constructs are the reliability coefficient, the standard error of measurement, and the confidence interval.

Reliability coefficients and correlational procedures.

In the case of a completely reliable test, the obtained score is equal to the true score. When an obtained score does not equal the true score, the difference is attributed to measurement error and regression to the mean. Since the true score cannot be known, the difference between two administrations of the same test are thought to be an estimate of the difference between an obtained score and a true score. For a given sample of examinees, the average of the product of the difference between the *z* score values of the two scores of all examinees produces a correlation coefficient that is referred to as the reliability coefficient.

Reliability coefficients are derived from correlational procedures. Correlation is a statistical procedure used to measure and describe the relationship between two variables or scores (Anastasi & Urbina, 1997; Gravetter & Wallnau, 2007). Correlational methods provide us with information regarding the direction and degree of the relationship between two scores. In terms of direction, correlations can be either positive or negative. Positive correlations indicate that the two scores or variables move in the same direction;

as one variable increases, so does the other. Negative correlations indicate an inverse relationship between scores; as one score increases, the other decreases.

Correlation coefficients can range from -1.00 to 1.00. A coefficient of .00 would indicate no relationship between scores, while a coefficient of 1.00 indicates a perfectly consistent relationship between scores (all scores are either positively or negatively related in an identical manner) (Anastasi & Urbina, 1997). The correlation coefficient can represent an index of reliability; a coefficient of .00 would indicate total unreliability, reflected in a lack of relationship between scores, while a coefficient of 1.00 would indicate perfect reliability, that is, a perfectly consistent relationship between scores. A test with a reliability coefficient of .80, for example, contains less measurement error and is more reliable than a test with a reliability coefficient of .40. For standardized-assessment instruments, test developers and researchers usually consider reliability coefficients of .90 to be excellent, while those in the .80's are good and those in the .70's are adequate; however, these descriptive classifications vary depending on the type and use of the scores being correlated (Anastasi & Urbina, 1997).

Standard error of measurement and confidence intervals.

The standard error of measurement (*SEM*) is a mathematical formula constructed around the reliability coefficient to quantify in a more specific manner the estimate of the amount of error associated with test scores (Psychological Corporation, 2004). The *SEM* represents the standard deviation of the distribution of scores around the hypothetical true score (Nunnally & Bernstein, 1994). A large *SEM* indicates a high level of error associated with observed score efforts to approximate the true score and therefore reflects poor reliability. A small *SEM* indicates a small amount of error associated with observed score efforts to approximate the true score and therefore reflects a high level of reliability.

Confidence intervals represent a range of score values based on the *SEM* within which an individual's true score is contained (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994). The size of the confidence interval varies based on the degree of certainty that the true score is actually contained within the confidence-interval score range. Confidence intervals are a way to represent the effect that unreliability has on score estimation. They illustrate the fact that error is present in all scores.

Procedures for Calculating Test-Retest and Internal Consistency Reliability

Test-retest and internal consistency reliability for the Wechsler Scales (Wechsler, 1997; 2002; 2003) are calculated using the Pearson Product-Moment and Spearman-Brown correlational formulas, respectively.

The Pearson Product-Moment correlational formula used to calculate stability coefficients is as follows (Nunnally & Bernstein, 1994):

$$r = \frac{(\mathrm{Zx})(\mathrm{Zy})}{n}$$

Lowercase *r* is the symbol for the correlation coefficient, and it is derived by dividing the sum of the product of the *z* scores by *n*. Note that, in this case, (Zx) is the time 1 score and (Zy) is the time 2 score.

With the exception of the processing speed subtests, whose reliability is represented using the stability coefficient only, internal consistency for the Wechsler

Scales subtests and composites was calculated using the split-half method. Application of the split-half procedure is not appropriate for speeded subtests because of their structure, which prevents the creation of two equivalent half-tests (Nunnally & Bernstein, 1994). For instance, consider the WISC-IV Coding subtest, which requires the child to use a grid to correctly match as many numbers and shapes as possible within a 2-minute time limit. Because the number of correct responses within the 2-minute time period will vary between children, it is not possible to adequately split the subtest into two equivalent halves.

The internal-consistency reliability coefficient is derived by correlating the total scores for the two half-tests. Unlike stability coefficients, internal-consistency coefficients are based on the whole normative sample. The Spearman-Brown formula recommended by Guilford (1954) and Nunnally and Bernstein (1994) is used because it corrects for the loss of items during the split-half procedure. When using Pearson's formula, loss of test items can lead to lower estimations of reliability. The Spearman-Brown procedure allows for prediction of what the reliability coefficients would be if entire subtests were used in the correlational formula. The Spearman-Brown correction formula is as follows (Nunnally & Bernstein, 1994):

$$r_{nn} = \frac{Kr_{tt}}{1 + (k - 1)r_{tt}}$$

The estimated reliability coefficient is r_{nn}, *k* is the number of items on the half-test divided by the number of items on the original test, and r_{tt} is Pearson's *r* before correction.

Procedures for Calculating Confidence Intervals

The first step in calculating confidence intervals is to establish the level of confidence based on the degree of certainty preferred. The test manuals for the most commonly used intelligence tests typically report confidence intervals for full scale and composite scores based on 90% and 95% confidence levels. Confidence intervals correspond to *z* scores and percentages of area under the normal curve. For example, the 95% confidence interval covers 95% of the normal curve and is associated with a *z* score of -1.96 and +1.96. When the confidence level is set at 95%, we can say that there is a 95% chance that the individual's true score falls within the interval of scores calculated. So, if the confidence interval associated with an IQ score of 100 is 96-105, classical test theory would have us conclude that the individual's true IQ score is somewhere between 96 and 105.

After the confidence level is selected, there are two procedures for calculating confidence intervals. Before describing these procedures, it is important to note that when subtest reliability estimates are high, which is typically the case for the Wechsler Scales, the confidence intervals established by both procedures show little difference (Psychological Corporation, 2004).

The formula for the first method is based on obtained scores and the *SEM* and is as follows (Sattler, 2001):

$$P\text{\% Confidence Interval} = \text{Observed Score} \pm Z_p(SEM)$$

P% is the desired confidence level, Z_p is the *z* score associated with the confidence level, and *SEM* is the standard error of measurement. The upper limit of the confidence interval is computed by adding the product $Z_p(SEM)$ to the observed score, while the lower limit is computed by subtracting the product from the observed score.

An alternative method is based on the estimated true score and the standard error of estimation (*SEE*) (Dudek,1979; Glutting, McDermott & Stanley, 1987). The *SEE* is the average standard deviation of true scores around an obtained score (Nunnally & Bernstein, 1994). It is calculated using the formula proposed by Stanley (1971):

$$SEE = SD\ (r_{xx})\ \sqrt{}1\text{-}r_{\mathrm{xx}}$$

SEE is the standard error of estimation, *SD* is the theoretical standard deviation of the composite or subtest, and r_{xx} is the subtest or composite reliability coefficient.

The estimated true score is calculated using the following formula (Sattler, 2001):

$$\mathrm{T} = r_{xx}\, x + (1 - r_{xx})\, X$$

T is the estimated true score, r_{xx} is the reliability of the test composite, *x* is the obtained score, and *X* is the mean of the test. The estimated true score is used in this formula because it is hypothesized to be closer to the mean of the test than an individual's observed test score. And when used with the *SEE,* the estimated true score is a correction for true-score regression to the mean (Psychological Corporation, 2004).

When confidence intervals are calculated using the estimated true score rather than the observed score, they can, in theory at least, provide a zone of expectation within which scores from re-administrations of the test are expected to fall. This is because, according to the concept of regression to the mean, individuals' performances with retesting should more closely approximate the mean of the scale and their true score. The hypothetical concept of regression to the mean states that when there is a less-than-perfect correlation between two administrations of the same test, extreme scores (i.e., scores farther from the group mean) tend to move toward the mean on the second administration (Gravetter & Wallnau, 2007). Of course, confidence intervals have predictive value only when reliability coefficients are high and *SEM*'s are low. With increased measurement error come wider confidence intervals and decreased ability to accurately predict retest performance.

Factors Affecting Reliability Estimates Based on Correlational Methods

Numerous factors associated with examinee performance patterns and the characteristics of examiners, examinees, the instrument being used, and the testing environment can affect reliability estimates calculated using correlational procedures. The following is a brief review of these factors.

Restriction of range.

Restriction of range refers to a situation in which the range of scores used to estimate the reliability of an instrument is not representative of the full range of scores in the population (Anastasi & Urbina, 1997; McCloskey, 1990). Restriction of range may lead to underestimations of reliability.

To illustrate, consider a group of students chosen as the standardization sample on a test measuring word knowledge. If the majority of these students earned scaled scores between 7 and 9, with the full range of possible scores being 1-16, we can assume that this is a restricted range; testing of additional students in the population will likely produce a number of scaled scores lower than 7 and higher than 9, making the test appear unreliable.

Outliers.

Similar to scores that are restricted in range, outliers also can have an impact on estimations of reliability. Outliers are individual scores that are substantially lower or higher than the scores obtained by the majority of the group. When using correlational methods to calculate test reliability based on the scores of a group of individuals, the presence of only one score that is substantially higher or lower than the others can lead to an underestimation of reliability.

Test length.

In general, the longer the test, the more reliable it is (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994). Consequently, in traditional psychometric orientations, many argue that global scores are more reliable, as they are comprised of many more items than are subtest or factor scores.

Guessing.

Guessing occurs when individuals respond arbitrarily to test items. Even when guessing results in correct answers, it introduces error into scores and can reduce estimates of reliability (Nunnally & Bernstein, 1994).

Variations within the testing situation.

Variations within the testing situation refers to any behaviors by examiners during testing that introduce error into testing procedures. Examples include examiner scoring errors and incorrect responses as a result of the examinee not understanding directions (Nunnally & Bernstein, 1994).

Limitations of Correlational Procedures When Used to Estimate Intelligence-Test Reliability

Use of correlational procedures to estimate intelligence-test reliability is well documented in the literature, including studies examining short-term test-retest reliability (Psychological Corporation, 2004) and long-term stability of IQ scores among different demographic subgroups (Canivez & Watkins, 1999) and learning-disabled populations (Kaye & Baron, 1987). However, there are several limitations to this statistical approach related to the quantity and quality of information it provides test consumers regarding test-retest score variability. These limitations are described next.

Use of a mean score to represent test score variability.

Close inspection of the formula for calculating Pearson's correlation, r = Sum (Zx)(Zy)/n, reveals that this formula actually provides an averaging of variability in test-retest scores. The question posed here is whether an average is the best quantitative method for representing variability in performance from time 1 to time 2. This descriptive statistic not only is sensitive to outliers but also, when used to describe large data sets, provides only limited information about the distribution of scores from which it was calculated.

Correlation and causation.

Correlation provides an estimate of the degree and direction of the relationship between two test scores but does not provide information regarding cause and effect. In other words, correlation coefficients can tell us that two test scores are related, or not related, but they do not tell us why. As an example, consider a situation in which the test-retest reliability coefficient for a test measuring vocabulary knowledge is .20, suggesting low reliability. This low reliability coefficient tells us that the students in the test-retest sample did not perform the same during both administrations of the test but gives no indication as to whether the variable performance was due to measurement error, practice effects, etc.

Inflation of error estimates.

Estimates of error derived from reliability coefficients tend to be inflated because any change in an individual's score from time 1 to time 2 that may be the result of increases in cognitive efficiency is added to the measurement error unless every student in the sample shows similar changes (Salvia & Ysseldyke, 2004). Stated another way, because reliability estimates are based on groups of individuals and not on individual scores, changes in scores as a result of practice effects or other cognitive factors as previously suggested are interpreted as measurement error.

Coefficient of determination.

Related to practice effects, another example of how classical test theory and the correlational approach may overestimate measurement error and underestimate changes in test performance due to changes in cognitive efficiency can be found in the concept of coefficient of determination. Calculating the coefficient of determination is another way to assess the amount of error attached to test-retest reliability coefficients. To calculate the coefficient of determination, one must simply square the correlation coefficient. The coefficient of determination provides an indication of the amount of variability in Score 2 that can be attributed to Score 1.

For example, if the test-retest correlation coefficient for the WISC-IV arithmetic subtest is .79, the coefficient of determination would be .62. This means that 62% of the variance in the retest score can be attributed to the original test score, while 38% of the variance, according to classical test theory, is attributed to error, even if the variability is due to meaningful differences in the use of cognitive capacities rather than to random fluctuations.

Alternatives to the Traditional Psychometric Theoretical Conception of Reliability

Item response (Rasch & Lord, 1960) and generalizability theory (Cronbach, Rajaratnam & Gleser, 1963), also known as G-theory, were developed as alternatives to the classical test theory model of reliability. The primary purpose of item response theory (IRT) and its derivatives (Rasch Model; Wright & Masters, 1982; Faceted Rasch Model; Linacre, 1989; Multidimensional Item Response Theory Models; McDonald, 1967) is to estimate the underlying theoretical trait that is presumed to contribute to an individual’s observed response on a measure (Suen & Lei, 2007). This is accomplished via the use of a probabilistic model of response, the logistic ogive model (Suen & Lei, 2007).

In contrast, the focus of G-theory (Cronbach, Rajaratnam & Gleser, 1963), like classical test theory, is the estimation of reliability for whole tests. However, there are some important theoretical differences between G-theory and classical test theory, most notably in their conceptualizations of the components of observed scores. As noted previously, in classical test theory an individual's observed score is hypothesized to consist of both true score and error variance. As stated in Suen and Lei (2007), in G-theory, the observed scores of all examinees on all items of the instrument are conceptualized as consisting of a universe score, µ, which is the theoretical mean of all the item scores in the universe of items, the deviation of the examinees' average item responses from µ, other deviations from µ, and random error.

Within the G-theory model, other deviations or sources of variance are referred to as facets, which may include different items, raters, times, or forms. Reliability estimates based on these different facets are calculated directly via the common analysis of variance statistical method (Suen & Lei, 2007). The advantages of G-theory are that it enables the consideration of multiple sources of error simultaneously and does not assume linearity, as is the case with traditional psychometric approaches. However, in circumstances where multiple sources of error are analyzed, very complex and lengthy statistical formulas are required to calculate reliability coefficients and *SEM*s. As is the case with traditional psychometric methods, G-theory does not consider the neuropsychological implications of exposing the brain to information that it can use to modify performance when presented with the same or similar information at a later time. The complexity of calculation methods and a reliance on hypothesized theoretical constructs, much like traditional psychometric theory, suggest that implementation of G-theory is likely to obscure rather than to clarify the issue of reliability for clinicians who use the Wechsler Scales on a regular basis.

Alternatives to Correlational Procedures for Estimating Test Reliability

Decision consistency models.

The use of decision-consistency models is well established in the literature on reliability estimations for criterion-referenced tests (Subkoviak, 1980; Traub & Rowley, 1980; Van Der Linden, 1980). Criterion-referenced instruments use a specific criterion to evaluate individual performance. For example, for a test measuring word knowledge, raw scores at or above 65 may be classified as passing scores while scores below 65 are failing.

In contrast to norm-referenced instruments where reliability estimates are based solely on test score variability from the first testing to subsequent testing, actual raw score variability between test and retest is less relevant when determining criterion-referenced test reliability (Traub & Rowley, 1980). What is important and what must be considered in a reliability assessment is the precision with which the instrument yields similar classifications according to the set criterion when repeated testing occurs. In the previous example, reliability could be assessed as the regularity with which individuals who were classified as passing on trial 1 also received a passing score on trial 2.

The basic procedures for applying a decision-consistency model are relatively simple and straightforward. Using data from two administrations of the same test, one can calculate the percentage of agreement for individuals who are classified the same on both test administrations according to the designated cut score. Decision-consistency models often use statistical procedures, such as chi square and coefficient kappa, to evaluate the degree of agreement between test and retest scores (Swaminathan, Hambleton, & Algina, 1974).

Use of Decision-Consistency Models with Norm-Referenced Assessments

Leach, Kaplan, Dymtro, Richards and Proulx (2000) utilized a decision-consistency model when calculating and presenting test-stability data for the Kaplan-Baycrest Neurocognitive Assessment (KBNA). The KBNA is an instrument commonly used by neuropsychologists to assess the integrity of cognitive functioning in the areas of attention, verbal fluency, spatial processing, reasoning and conceptual shifting, and immediate and delayed memory. In addition to reliability coefficients calculated using traditional correlational procedures, the KBNA authors present decision-consistency percentages indicating the percentage of the standardization sample whose classification range (Below Average, Equivocal, Average) on each subtest did not change from test to retest.

In another variation of a decision-consistency model, McCloskey (1990) compared an agreement grid to traditional correlational procedures when estimating the interrater reliability of two early-childhood-behavior rating scales. Design of the grid involved the calculation of three agreement percentages: (1) an identical ratings percentage showing the percentage of exact agreement between first and second rating; (2) an increased ratings percentage showing the percentage of increased ratings from time 1 to time 2; and (3) a decreased ratings percentage indicating the percentage of decreased ratings from the first to second rating.

Pearson's correlation coefficients for both scales were calculated by correlating time 1 and time 2 ratings. The correlation coefficient for the first scale was .59, suggesting poor consistency between time 1 and time 2 ratings. However, the agreement grid indicated a much higher level of decision agreement as evidenced by the overall 81.6 identical ratings percentage between the first and second ratings. McCloskey (1990) attributed the large discrepancy between the correlation coefficient and identical ratings percentage to a restricted range of ratings, which negatively influences the former but not the latter.

For the second rating scale, both the correlation coefficient (.84) and identical ratings percentage (76%) between the first and second ratings were high. The high percentage of increased ratings on the second scale (21.5%) may have been the result of expectancy effects or the tendency of raters to provide higher ratings over time in the absence of actual increases in target behavior (McCloskey, 1990).

McCloskey (1990) reported several advantages of an agreement grid over correlational procedures when attempting to establish the reliability of behavior rating scales. Among these is the agreement grid's ability to yield accurate information about degree of test-retest score agreement regardless of the score distribution. McCloskey also suggested that agreement grids can provide valuable information about the nature of test-retest score disagreements and facilitate a better understanding of the extent to which expectancy effects accounted for increases in ratings from time 1 to time 2.

The modified decision-consistency model illustrated by McCloskey (1990) could be adapted to examine the degree of consistency between WISC-IV subtest scores in a test-retest condition. Values for the degree of negative change, no change, and positive change could be calculated and compared with hypothetical results based on traditional psychometric theory, such as true-score estimates, and/or based on the neuropsychological literature related to practice effects.

A modified decision-consistency model approach to the analysis of test reliability has a number of potential advantages over traditional psychometric methods, including a more complete view of the variability of scores from time 1 to time 2 that does not mask important patterns in the movement of scores.

Conclusion

Despite its demonstrated value in diagnosis and intervention planning, subtest-level interpretation of intelligence tests continues to be criticized on the basis that subtests have poor reliability and poor interpretative power. However, it may be that perceptions of subtest instability are the result of misguided conceptualizations of and poor methods for representing test reliability and are not related to the actual properties of subtests.

The current psychometric model of test reliability assumes that intelligence tests measure specific cognitive constructs that are stable over time and, as such, that individual performance should not vary between test and retest. In cases where test-retest performance results in improvement that is not consistent across all individuals in the sample, it is assumed that this inconsistency is the result of measurement error and not systematic variance from sources other than those that are the focus of the assessment.

Viewing all forms of variance in performance from time 1 to time 2 as undesired measurement error that reduces the reliability of the test is not consistent with the neuropsychological literature base indicating that increases in performance are due to improved cognitive efficiency, commonly referred to as practice effects, that contribute in an expected and meaningful way to increases in performance on repeated measures of cognitive functioning (Kaufman, 2003; Goldberg, 2001; Matarazzo et al., 1980; Catron & Thompson, 1979). From a neuropsychological perspective, these increases in performance can be explained as resulting from changes in brain functioning that promote the learning of novel tasks and more efficient execution of familiar tasks. Rather than eschew these brain-state changes as measurement error that detracts from a test's usefulness and reliability as a source of information about examinee performance, test developers should develop and employ methods that enable clinicians to recognize and quantify expected changes in task performance in a manner that is meaningful and clinically interpretable.

Decision-consistency models offer a basis for the development of methods for conducting more meaningful reliability analyses that may prove superior to traditional psychometric methods in terms of clinical utility. A variation of the basic decision-consistency model was proposed by McCloskey (1990). This method involves the classification of the difference between time 1 and time 2 testing as negative change, no change, or positive change and has the advantage of enabling the clinician to simultaneously view and understand how, and the extent to which, scores do or do not change on repeated administrations.

Statement of the Problem

Given the perceived inadequacies of the traditional psychometric model to offer realistic, clinically meaningful information about variation in test performance based on repeated administration of subtest tasks, the present study sought to offer an alternate means of understanding and interpreting the concept of subtest reliability using a variation of a decision-consistency model incorporating neuropsychologically based knowledge to establish expected WISC-IV 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 model fit WISC-IV subtest test-retest data? Extrapolating from the findings in the literature, the following patterns of performance variation from time 1 testing to time 2 testing are hypothesized:

Hypothesis 1A: For a majority of cases, subtests that primarily involve retrieval from long-term storage or association with stored knowledge (Vocabulary, Information, Word Reasoning, Similarities) will 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 showing change will be biased toward a progression effect rather than a regression effect; that is, increases will outnumber decreases even in situations where regression to the mean would predict performance decreases or no change.

Hypothesis 1B: For a majority of cases, subtests that primarily involve the initial registration of and manipulation of verbal information in mind (Digit Span Forward and Backward, Digit Span, Letter-Number Sequencing, Arithmetic) will 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) will distribute relatively equally around a central tendency of no change with relatively fewer but equal numbers of cases demonstrating positive and negative change both above and below the mean. Tasks that primarily involve the initial registration and manipulation of verbal information that is more contextual and meaningful (Arithmetic and, to a lesser degree, Letter-Number Sequencing) will demonstrate a pattern of performance closer to that of tasks involving retrieval from long-term storage.

Hypothesis 1C: For a majority of cases, subtests that primarily involve novel problem solving (Matrix Reasoning, Picture Concepts, Block Design, Picture Concepts) or processing speed applied to simple but relatively novel tasks (Coding, Symbol Search, Cancellation) will yield score differences that reflect positive changes in performance, reflecting a greater progression than regression effect. Score decreases will be similar in magnitude both in cases where time 1 scores were above the mean and in cases where time 1 scores were below the mean due to the 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.

Hypothesis 1D: When considered in total, the test-retest results will 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 conception of reliability.

Question 2: Does a neuropsychologically oriented performance-consistency method offer any possible advantages over traditional psychometric methods in the type of information it provides test consumers regarding WISC-IV subtest performance patterns?

Chapter 2

Method

Participants

This study utilized the archival data set from the WISC-IV standardization sample used to calculate the test-retest reliability for this instrument. As reported in the *WISC-IV Integrated Technical and Interpretative Manual* (Psychological Corporation, 2004), the participants included 243 children (52.3% female & 47.7% male) between the ages of 6 and 16 years selected to be representative of the total WISC-IV standardization sample. Each participant was assessed on two occasions (mean test-retest interval 32 days) using all 15 subtests of the WISC-IV. 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 as follows: 4.9%, 0-8 years; 9.1%, 9-11 years; 25.9%, 12 years; 36.2%, 13-15 years; and 23.9%, greater than or equal to 16 years (Psychological Corporation, 2004).

Measures

The WISC-IV 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 *WISC-IV Integrated Technical and Interpretative Manual* (Psychological Corporation, 2004) provides detailed information regarding the instrument's validity and reliability. The WISC-IV has been shown to demonstrate adequate content, criterion-related, and construct validities. As reported in the *WISC-IV Integrated Technical and*

Interpretative Manual (Psychological Corporation, 2004), Tables 3 and 4 show the average *SEM*, reliability coefficients, and corrected stability coefficients for the subtests, composite scales, and process scores for the total sample.

Table 3

Average Reliability Coefficients, SEM, and Corrected Stability Coefficients for WISC-IV Subtests for the Total Sample

Subtest	r_{xx}	*SEM*	*r*
Block Design	.86	1.13	.82
Similarities	.86	1.13	.86
Digit Span	.87	1.07	.83
Picture Concepts	.82	1.29	.76
Coding	.85	1.20	.84
Vocabulary	.89	1.00	.92
Letter-Number Sequencing	.90	.97	.83
Matrix Reasoning	.89	.99	.85
Comprehension	.81	1.31	.82
Symbol Search	.79	1.36	.80
Picture Completion	.84	1.20	.84
Cancellation	.79	1.38	.79
Information	.86	1.16	.89
Arithmetic	.88	1.05	.79
Word Reasoning	.80	1.34	.82

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 4

Average Reliability Coefficients, SEM, and Corrected Stability Coefficients for WISC-IV Process Scores and Composite Scales for the Total Sample

Process Score	r_{xx}	*SEM*	r^a
Digit Span Forward	.83	1.24	.76
Digit Span Backward	.80	1.37	.74
Cancellation Random	.70	1.66	.72
Cancellation Structured	.75	1.51	.76
Verbal Comprehension	.94	3.78	.93
Perceptual Reasoning	.92	4.15	.89
Working Memory	.92	4.27	.89
Processing Speed	.88	5.21	.86
FSIQ	.97	2.68	.93

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 *SEM* is lowest for Letter-Number Sequencing (.97) and highest for Cancellation (1.38). For the composite scales and process scores, *SEM* is lowest for FSIQ (2.68) and highest for PSI (5.21) and Cancellation Random (1.66). Test-retest coefficients for the total sample ranged from .93 for the FSIQ and VCI to .86 for the PSI. Vocabulary had the highest test-retest reliability estimate (.92), and Picture Concepts had the lowest (.76).

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 243 cases from the standardization test-retest reliability study, the following procedures were carried out to complete the analyses:

1. For each WISC-IV subtest and selected process scores, 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 WISC-IV subtest and selected process scores, a Predicted Time 2 score was calculated using the following formula:

$$x2 = X2 + (x1 - X1) * rx1x2 * (sdx2 / sdx1)$$

where $x2$ = Time 2 score

$x1$ = Time 1 score

$X1$ = Mean of Time 1 scores (set at 10 for all subtests)

$rx1x2$ = the correlation between Time 1 and Time 2 scores

$sd1$ = the standard deviation of Time 1 scores (set at 3) sd2 = 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 since the term (sd2/sd1) is equal to 1 because 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 WISC-IV subtest and selected process scores, 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 WISC-IV subtest and selected process scores, frequency distributions were obtained for the Actual Difference scores and the Predicted Difference scores and tabled together for comparison and analysis.
5. For each WISC-IV subtest and selected process scores, the sample of 243 cases was divided into two groups to test the traditional psychometric conception of regression to the mean that was incorporated in the formula a Predicted Time 2 score.
    a. The LTE group consisted of all cases where the time 1 score was less than 10. This group comprised the cases that would be 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 the effect of regression to the mean. Also included

with this group were cases where the time 1 score was 10 and the time 2 score was less than 10 for reasons stated later.

b. The GTE group consisted of all cases where the time 1 score was greater than 10. This group comprised the cases that would be 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 the effect of regression to the mean. Also included with this group were cases where the time 1 score was 10 and the time 2 score was greater than 10 for reasons stated later.

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 T2-T1 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. Cases earning time 1 scores of 10 that reflected a negative Actual Difference score were included in the LTE group because they were expected to remain the same rather than to decrease, and cases earning a time 1 score of 10 that reflected a positive Actual Difference score were included in the GTE group because they were expected to remain the same rather than to increase. Only the cases earning time 1 scores of 10 that reflected no change in Actual Difference score were eliminated from the analysis. In all cases, the number of cases eliminated from the analysis was typically less than 10% of the total sample. This case assignment procedure represented a bias based on time 2 scores that ultimately was in favor of the traditional psychometric model, as other means of including or excluding the cases earning scores of 10 at time 1 would have further increased the proportions of cases not conforming to the expected no change/regression to the mean performance pattern.

6. For each WISC-IV subtest and selected process scores, Predicted Difference scores were assigned to one of three score-change categories:
   Negative Change (-), No Change (0), and Positive Change (+).
7. For each WISC-IV subtest and selected process scores, Actual Difference scores were assigned to one of three score-change categories: Negative Change (-), No Change (0), and Positive Change (+).
8. For each WISC-IV subtest and selected process scores, a 2 x 3 cross-tabulation table was generated indicating frequency counts of the three score-change categories for Actual Difference scores and Predicted Difference scores for the cases assigned to the LTE group. The frequencies in the 2 x 3 table were subjected to a chi-square analysis to determine goodness of fit between the Actual and the Predicted Difference proportions.

9. For each WISC-IV subtest and selected process scores, a 2 x 3 cross-tabulation table was generated indicating frequency counts of the three score-change categories for Actual Difference scores and Predicted Difference scores for the cases assigned to the GTE group. The frequencies in the 2 x 3 table were subjected to a chi-square analysis to determine goodness of fit between the Actual and the Predicted Difference proportions.
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 in order for a valid analysis to be completed. In situations where the score-change category count was 0 for the Predicted Difference score, five cases were removed from the No Change category and placed in the category with the 0 count, thereby enabling the completion of all chi-square analysis. 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, leading to a nonsignificant chi-square value.

Chapter 3

Results

Cross-tabulation analyses were conducted to determine the frequency of score differences between time 2 and time 1 administration for each WISC-IV subtest and select process scores. The Actual Difference scores were then compared with the Predicted Difference scores calculated using the regression model described in Chapter 2.

Table 5 shows the frequency distributions for the Actual and Predicted Differences between time 2 and time 1 performance on the WISC-IV Verbal subtests.

Table 5

Frequency Distributions of Actual and Predicted T2 – T1 Differences for Each WISC-IV Verbal Subtest

Subtest (n)	Difference														
Vocabulary (n = 242)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual					2	18	46	81	63	19	8	5			
Predicted								239	3						
Information (n = 243)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual				4	2	16	27	86	60	31	11	4	2		
Predicted							9	227	7						
Similarities (n =239)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual				1	4	14	32	67	60	32	20	8		1	
Predicted							24	191	24						
Comprehension (n = 234)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual			4	2	10	18	35	61	46	34	17	5	2		
Predicted						1	34	167	32						
Word Reasoning (n = 243)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual	1				6	18	31	49	56	44	25	8	2	3	
Predicted							43	167	33						

For all Verbal subtests with the exception of Comprehension, the Predicted Difference scores did not exceed -1 or +1. Across all Verbal subtests, a higher frequency of students demonstrated no change or positive scaled-score change versus negative scaled-score change. With the exception of the Word Reasoning subtest, where the largest number of students (n = 56) demonstrated a scaled-score increase of +1, the performance of the majority of examinees did not change from time 1 to time 2. For all subtests, the retest performance of the majority of examinees fell between -1 and 1, suggesting that this score band may prove useful in predicting WISC-IV Verbal retest performance using the alternative reliability model presented here. Word Reasoning produced the most test-retest score variability, with three examinees demonstrating score improvement of +6 and one examinee showing a decrease of -7 scaled-score points. A fairly large discrepancy between Actual and Predicted score frequencies was evident for all subtests. The only exception was Comprehension, where the Actual frequency of examinees showing a difference of -1 varied from the prediction model by only 1 (n = 35 versus 34). Most notable in the table is the prominence of Actual Difference increases over decreases that are more consistent with a neuropsychologically based performance model than with the traditional psychometric model.

Table 6 shows the frequency distributions for the Actual and Predicted Differences between time 2 and time 1 performance on the WISC-IV Working Memory subtests.

Table 6

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

Subtest (n)	Difference													
Digit Span (n = 241)														
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual				1	9	16	34	68	56	28	17	8	2	2	
Predicted						2	42	149	48						
Digit Span Forward (n = 243)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual			3	5	12	21	21	73	42	29	16	14	6		1
Predicted						3	39	148	53						
Digit Span Backward (n = 237)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual				10	8	28	23	74	22	35	14	15	5	2	1
Predicted						6	65	96	66	4					
Letter-Number Sequencing (n = 235)															
	-7-9	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual	1			1	6	21	42	65	46	21	18	8	6		
Predicted							41	164	29	1					
Arithmetic (n = 133)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual				10	8	28	23	74	22	35	14	15	5	2	1
Predicted						6	65	96	66	4					

For all Working Memory subtests, the Predicted Difference scores did not exceed -2 or 2. However, a large discrepancy between Actual and Predicted score frequencies was evident for all subtests. The only exception was Letter-Number Sequencing, where the Actual frequency of examinees showing a difference of -1 varied from the prediction model by only 1 (n = 42 versus 41). A greater number of examinees demonstrated no change or positive scaled-score change on the Working Memory subtests. Across all subtests, the largest number of examinees demonstrated no scaled-score change between time 1 and time 2. Performance on all Working Memory subtests was characterized by higher frequencies of Actual Difference increases over decreases that are more consistent with a neuropsychologically based performance model than with the traditional psychometric model.

Table 7 shows the frequency distributions for the Actual and Predicted Differences between time 2 and time 1 performance on the WISC-IV Perceptual Reasoning subtests.

Table 7

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

Subtest (n)	Difference														
Block Design (n = 240)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual					3	7	24	63	54	34	25	20	8	1	1
Predicted						1	51	143	44	1					
Matrix Reasoning (n = 239)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual				5	8	9	29	72	47	32	23	8	5	1	
Predicted							26	196	17						
Picture Concepts (n = 234)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual		1		2	9	24	27	39	46	37	21	18	7	2	1
Predicted						1	41	149	41	2					
Picture Completion (n = 243)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual						4	13	46	49	57	37	23	6	5	3
Predicted							26	191	26						

For all Perceptual Reasoning subtests, the Predicted Difference scores did not exceed -2 or 2. As with the Verbal and Working Memory subtests, though, the Predicted scores based on the regression model had limited predictive validity. Only on Matrix Reasoning did the Predicted regression frequency closely approximate the Actual frequency (n = 26 versus 29). Across all Perceptual Reasoning subtests, a higher frequency of students demonstrated no change or positive scaled-score change versus negative scaled-score change. On both Block Design and Matrix Reasoning, the highest retest score frequency was 0, n = 63 and 72, respectively. For Picture Concepts, the largest number of examinees demonstrated retest scaled-score increases of +1 (n = 46), while for Picture Completion, the most common rate of improvement was +2 (n = 57). Overall, Picture Completion showed the highest rates of score variability and score improvement. Consistent with the neuropsychologically based performance model, across all subtests there were higher frequencies of Actual Difference increases over decreases.

Table 8 shows the frequency distributions for the Actual and Predicted Differences between time 2 and time 1 performance for the WISC-IV Processing Speed subtests and process scores.

Table 8

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

Subtest (n)	Difference														
Coding (n = 231)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual						11	25	39	46	51	27	20	7	2	3
Predicted							35	178	18						
Symbol Search (n = 233)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual		1		4	6	11	24	43	46	45	24	14	10	3	2
Predicted							49	155	28	1					
Cancellation (n =234)															
	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7-9
Actual			2	2	5	14	27	34	50	41	30	13	11	4	1
Predicted							50	145	35	4					

For all Processing Speed subtests, the Predicted Difference scores did not exceed -1 or 2. However, for all subtests, there was a large discrepancy between the frequencies predicted by the regression model and the Actual Difference frequencies. Across all Processing Speed subtests, a greater number of examinees demonstrated no scaled-score change or positive scaled-score change versus negative scaled-score change. Unlike the subtests contained in the other WISC-IV indexes, all Processing Speed subtests showed higher frequencies of positive versus negative scaled-score change. The Processing Speed subtests also yielded the highest rates of score improvement and highest frequencies of significant test-retest score variability; for each subtest, 12-15 participants demonstrated score improvement of 5 or more scaled-score points. As with the subtests in the other domains, performance on the Processing Speed subtests was characterized by higher frequencies of Actual Difference increases over decreases.

Tables 5 through 8 present Actual and Predicted score frequency data for the total sample without clear indication of the values of the time 1 scores. 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 in relation to the mean of the scale. The regression model will predict score increases for more extreme time 1 scores below the mean and score decreases for more extreme time 1 scores above the mean.

Tables 9 through 12 show the percentages of time 2-time 1 score differences that reflect regression (-), progression (+), or no change (0). Also provided are the percentages of time 2-time 1 score differences that the regression model predicts would result in score regression, progression, or no change. The data are grouped by time 1 standard score ranges and time 2 score change categories using the procedures outlined in Chapter 2.

Table 9

Actual and Predicted Difference Score Percentages by Score Change Category for Each WISC-IV Verbal Subtest for the LTE and GTE Groups

Subtest (n)	T1 Scaled Scores LTE 10				T1 Scaled Scores GTE 10			Subtest (n)
Vocabulary (n = 102)	-	0	+		-	0	+	Vocabulary (n = 125)
Predicted	0%	99%	1%		0%	100%	0%	Predicted
Actual	31%	28%	40%		27%	30%	43%	Actual
Information (n = 105)	-	0	+		-	0	+	Information (n = 119)
Predicted	0%	93%	7%		9%	91%	0%	Predicted
Actual	18%	27%	55%		25%	33%	42%	Actual
Similarities (n = 95)	-	0	+		-	0	+	Similarities (n = 131)
Predicted	0%	74%	26%		21%	79%	0%	Predicted
Actual	18%	21%	61%		26%	26%	48%	Actual
Comprehension (n = 101)	-	0	+		-	0	+	Comprehension (n = 121)
Predicted	0%	66%	34%		34%	66%	0%	Predicted
Actual	24%	24%	53%		37%	21%	42%	Actual
Word Reasoning (n = 103)	-	0	+		-	0	+	Word Reasoning (n = 132)
Predicted	0%	67%	33%		40%	60%	0%	Predicted
Actual	14%	22%	67%		32%	14%	54%	Actual

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 between the two groups using the procedure described in Chapter 2. Group n counts vary because of deletion of cases based on the case assignment procedures described in Chapter 2.

For all Verbal subtests, both the LTE and GTE groups demonstrated higher percentages of Actual score progression versus Actual score regression and lower percentages of no change than predicted. The percentages of Actual score progression were higher for the LTE than the GTE group for all subtests except Vocabulary (40% versus 43%). Across all subtests, both groups demonstrated higher percentages of Actual score progression than were predicted. The LTE group demonstrated higher rates of Actual versus Predicted score regression for all Verbal subtests. For the GTE group, percentages of Predicted score regression were generally lower than percentages of Actual score regression. The one exception was Word Reasoning, where the model predicted 40% regression and the actual was 32%. The GTE group demonstrated a higher rate of Actual score regression than that of the LTE group for all subtests except Vocabulary (31% versus 27%). The LTE group demonstrated higher percentages of Actual score progression than those of the GTE group for all subtests except Vocabulary (40% versus 43%),

Table 10

Actual and Predicted Difference Score Percentages by Score Change Category for Each WISC-IV Working Memory Subtest and Process Score for the LTE and GTE Groups

Subtest (n)	T1 Scaled Scores LTE 10				T1 Scaled Scores GTE 10			Subtest (n)
Digit Span (n = 115)	-	0	+		-	0	+	Digit Span (n = 117)
Predicted	0%	53%	47%		45%	55%	0%	Predicted
Actual	22%	27%	51%		30%	24%	46%	Actual
DS Forward (n = 102)	-	0	+		-	0	+	DS Forward (n = 136)
Predicted	0%	45%	55%		34%	66%	0%	Predicted
Actual	13%	32%	57%		36%	26%	38%	Actual
DS Backward (n = 104)	-	0	+		-	0	+	DS Backward (n = 121)
Predicted	0%	28%	72%		66%	34%	0%	Predicted
Actual	20%	31%	49%		40%	25%	35%	Actual
Letter Number (n = 89)	-	0	+		-	0	+	Letter Number (n = 136)
Predicted	0%	63%	37%		36%	64%	0%	Predicted
Actual	25%	24%	51%		36%	25%	39%	Actual
Arithmetic (n = 51)	-	0	+		-	0	+	Arithmetic (n = 65)
Predicted	0%	52%	48%		33%	67%	0%	Predicted
Actual	16%	19%	65%		35%	28%	37%	Actual

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 between the two groups using the procedure described in Chapter 2. Group n counts vary because of deletion of cases based on the case assignment procedures described in Chapter 2.

The LTE group demonstrated higher percentages of Actual score progression versus regression across all subtests. With the exception of Digit Span Backwards, where the Actual progression-regression percentages were 35% and 40%, respectively, the GTE group also demonstrated higher percentages of progression versus regression. However, it should be noted that, in contrast to subtests in other domains, the Working Memory subtests produced progression and regression percentages for the GTE group that were close in value. Likewise, with the exception of Digit Span Backwards, the GTE Predicted regression percentages closely approximated the Actual regression percentages. For Letter-Number Sequencing, the Actual and Predicted regression percentages were equivalent. Actual progression and regression percentages were higher than predicted for all subtests in the LTE group. Overall, the LTE group showed higher percentages of score progression and the GTE group demonstrated greater percentages of regression.

Table 11

Actual and Predicted Difference Score Percentages by Score Change Category for Each WISC-IV Perceptual Reasoning Subtest for the LTE and GTE Groups

Subtest (n)	T1 Scaled Scores LTE 10				T1 Scaled Scores GTE 10			Subtest (n)
Block Design (n = 100)	-	0	+		-	0	+	Block Design (n = 131)
Predicted	0%	55%	45%		51%	49%	0%	Predicted
Actual	10%	20%	70%		18%	26%	56%	Actual
Matrix Reasoning (n = 107)	-	0	+		-	0	+	Matrix Reasoning (n = 118)
Predicted	0%	47%	53%		27%	73%	0%	Predicted
Actual	20%	28%	52%		25%	24%	51%	Actual
Picture Concepts (n = 91)	-	0	+		-	0	+	Picture Concepts (n = 135)
Predicted	0%	47%	53%		40%	60%	0%	Predicted
Actual	24%	13%	63%		30%	14%	56%	Actual
Picture Completion (n = 95)	-	0	+		-	0	+	Picture Completion (n = 142)
Predicted	0%	73%	27%		21%	79%	0%	Predicted
Actual	3%	17%	80%		10%	17%	73%	Actual

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 between the two groups using the procedure described in Chapter 2. Group n counts vary because of deletion of cases based on the case assignment procedures described in Chapter 2.

Both the LTE and GTE groups demonstrated higher percentages of Actual score progression than regression for all subtests. The LTE group demonstrated higher percentages of Actual score progression than the GTE group across all subtests. Actual regression percentages were higher for the GTE group than for the LTE group. The Predicted percentages of progression were lower than the Actual percentages for all subtests for both groups. The LTE group showed higher percentages of Actual versus Predicted regression, while the GTE group demonstrated lower percentages of regression than predicted. Picture Completion showed the highest percentages of score progression, with 80% of the LTE group and 73% of the GTE group demonstrating improvement. For both groups, the percentages of Predicted no score change were much higher than the percentages of Actual score change. For all subtests, the percentages of score progression were higher for the LTE group than for the GTE group. However, the GTE group showed higher percentages of regression for all subtests.

Table 12

Actual and Predicted Difference Score Percentages by Score Change Category for Each WISC-IV Processing Speed Subtest for the LTE and GTE Groups

Subtest (n)	T1 Scaled Scores LTE 10				T1 Scaled Scores GTE 10			Subtest (n)
Coding (n = 94)	-	0	+		-	0	+	Coding (n = 131)
Predicted	0%	79%	21%		31%	69%	0%	Predicted
Actual	17%	15%	68%		15%	15%	70%	Actual
Symbol Search (n = 80)	-	0	+		-	0	+	Symbol Search (n = 147)
Predicted	0%	63%	37%		44%	56%	0%	Predicted
Actual	16%	18%	66%		22%	16%	62%	Actual
Cancellation (n = 103)	-	0	+		-	0	+	Cancellation (n = 128)
Predicted	0%	61%	39%		44%	56%	0%	Predicted
Actual	14%	11%	75%		27%	16%	57%	Actual

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 between the two groups using the procedure described in Chapter 2. Group n counts vary because of deletion of cases based on the case assignment procedures described in Chapter 2.

For all Processing Speed subtests, both the LTE and GTE groups demonstrated higher percentages of Actual score progression versus Actual score regression and lower percentages of no change than predicted. For the LTE group, percentages of progression and regression for all subtests were higher than predicted. Actual regression percentages for all subtests for the GTE group were lower than predicted. Progression percentages for all subtests in the GTE group were higher than predicted. Overall, with the exception of Cancellation (75% LTE versus 57% GTE), percentages of Actual progression for both groups were roughly equivalent. Percentages of Actual score regression were higher for the GTE group than those for the LTE group for all subtests except Coding (15% versus 17%).

Chi-square analyses were conducted to determine the goodness of fit between Actual and Predicted time 2 – time 1 score differences for all subtests and process scores. As indicated in Table 13, the chi-square analyses yielded statistically significant results beyond the .01 level for all subtests across the LTE and GTE groups.

Table 13

Chi-Square Analysis of Actual (Observed) and Predicted (Expected) T2 – T1 Differences for Each WISC-IV Subtest Grouped by Time 1 Standard Score Ranges and Time 2 Regression Categories

	T1 Scaled Scores LTE 10				T1 Scaled Scores GTE 10		
Subtest (n)	$\chi 2$	*Df*	*p*		$\chi 2$	*Df*	*p*
Vocabulary (n = 102)	448.14	2	<.01	Vocabulary (n = 125)	701.30	2	<.01
Information (n = 105)	456.20	2	<.01	Information (n = 119)	495.48	2	<.01
Similarities (n = 95)	109.27	2	<.01	Similarities (n = 131)	722.3	2	<.01
Comprehension (n = 101)	110.98	2	<.01	Comprehension (n = 121)	464.77	2	<.01
Word Reasoning (n = 103)	79.67	2	<.01	Word Reasoning (n = 132)	921.52	2	<.01
Block Design (n = 100)	74.93	2	<.01	Block Design (n = 131)	962.12	2	<.01
Matrix Reasoning (n = 107)	176.25	2	<.01	Matrix Reasoning (n = 118)	645.62	2	<.01
Picture Concepts (n = 91)	84.70	2	<.01	Picture Concepts (n = 135)	1034.12	2	<.01
Picture Completion (n = 95)	132.95	2	<.01	Picture Completion (n = 142)	2033.92	2	<.01
Digit Span (n = 115)	98.02	2	<.01	Digit Span (n = 117)	505.57	2	<.01
DS Forward (n = 102)	16.37	2	<.01	DS Forward (n = 136)	455.82	2	<.01
DS Backward (n = 104)	56.66	2	<.01	DS Backward (n = 121)	301.25	2	<.01
Letter Number (n = 89)	86.5	2	<.01	Letter Number (n = 136)	497.20	2	<.01
Arithmetic (n = 51)	8.68	2	<.01	Arithmetic (n = 65)	82.77	2	<.01
Coding (n = 94)	187.51	2	<.01	Coding (n = 131)	1577.19	2	<.01
Symbol Search (n = 80)	54.92	2	<.01	Symbol Search (n = 147)	1537.11	2	<.01
Cancellation (n = 103)	96.07	2	<01	Cancellation (n = 128)	967.77	2	<.01

Chapter 4

Discussion

Summary of Results

The present study was designed with two goals in mind. The first was to determine if a neuropsychologically based performance model better fit WISC-IV subtest test-retest data than does a traditional psychometric model. The second goal was to examine the utility of this model compared to traditional psychometric procedures in terms of the type of information it provides test consumers about WISC-IV subtest performance patterns.

Results related to the first goal indicated that, regardless of how the data were grouped (total sample versus time 1 standard score and time 2 regression categories of LTE and GTE), the performance consistency model (McCloskey, 1990) better represented the actual pattern of WISC-IV subtest test-retest differences than did the psychometric model (Anastasi & Urbina, 1997). Cross-tabulation analyses for the total combined sample revealed significant differences between the obtained and predicted score-difference frequencies for most performance levels across all subtests.

As expected, the range of predicted score-difference frequencies for all subtests did not extend beyond the -2 to 2 score band. For most subtests, there was a greater frequency of positive versus negative changes above the -2 to 2 score-difference band, regardless of whether the initial scores were above or below the mean. Likewise, greater numbers of extreme test-retest score differences (beyond 4 points) were observed for positive versus negative scaled-score differences. These findings indicate that, as predicted with a neuropsychologically based performance model, examinees were more likely to show test-retest score increases than decreases, and, in some cases, the frequency and number of performance improvements were exceptionally large.

Similar trends in the data were observed when the sample was grouped by time 1 standard score categories of LTE and GTE. As a neuropsychological model would predict (McCloskey, 2009), chi-square analyses indicated a poor fit between actual and predicted score differences for all subtests for both groups. These findings have two implications: (1) that there is no difference in the predictive power of the psychometric model for time 1 scores that are above or below the mean and (2) that WISC-IV subtest scores do not show the patterns of regression to the mean predicted by the psychometric model.

Except in cases where the prediction was no score change from time 1 to time 2, the psychometric model predicted score decreases (regression downward toward the mean) at time 2 when time 1 scores were above the mean and score increases (progression in the form of regression upward toward the mean) at time 2 when time 1 scores were below the mean. For the LTE group, this assumption generally held true. A higher percentage of Actual score increases versus Actual score decreases was demonstrated for all subtests and process scores. However, for all subtests, with the exception of Matrix Reasoning, the percentage of Actual score increases was greater than the percentage of Predicted score increases, suggesting a progression effect beyond what was expected based on regression toward the mean. Likewise, in all cases, there was a higher percentage of Actual score progression than no change even when no change was predicted.

For the GTE group, the regression model predicted that most examinees would demonstrate no test-retest score change or negative change. The GTE group did demonstrate some score regression to the mean. However, for all subtests with the exception of Digit Span Backward (40% regression toward the mean versus 35% progression away from the mean), the percentage of actual score progression away from the mean was higher than the percentage of actual score regression toward the mean. Likewise, for all subtests, there was a higher percentage of actual score progression versus actual no change and a much higher percentage of actual versus predicted progression away from the mean.

Additional evidence for the proposition that the movement of scores toward the mean in the LTE group was due more to a true progression effect based on brain-state changes that enabled more efficient performance than to a regression to the mean effect based on a higher probability of a reduced effect of measurement error is the fact that the score increases of the LTE group were even greater.

A secondary goal of this study was to examine the effectiveness of the performance consistency model in conveying information about WISC-IV subtest reliability beyond that which is provided by traditional psychometric procedures. Traditional methods for estimating test reliability are based on the assumption that intelligence is a static trait of which individuals possess a “true” or actual amount (Hale & Fiorello, 2004). Subtests assessing intellectual functioning, if perfectly reliable in measuring one or more facets of this static trait, will show no variability during test-retest performance. Also, in cases where test-retest score variation does occur, these fluctuations in performance are attributed to the effects of randomly distributed measurement error.

Applying this methodology to the present study, we should have found essentially no differences between time 1 and time 2 scores, or only minor fluctuations based on the magnitude of the correlation coefficients derived from the data. Also, if measurement error contributed to score variation, and it was the only contributor, it should have been randomly distributed, leading to equal instances of score regression and progression. However, results of the present study indicated overwhelmingly a pattern of retest score progression for all of the WISC-IV subtests and process scores. This finding suggests that conceptions of reliability based on traditional psychometric theory (Anastasi & Urbina, 2007) are not a good fit with the manner in which individuals actually perform on re-administrations of the same tasks.

The alternative presented here is a neuropsychologically based performance model that better captures the dynamic nature of brain-behavior relationships and the manner in which individuals use their brains when they engage tasks (Goldberg, 2001). Within this model, increases in retest performance are conceptualized as resulting from changes in neural activity that promotes the learning of novel tasks and the more efficient execution of previously learned skills (Goldberg & Costa, 1981; Raichle et al., 1994). Error is conceptualized as lack of consistency in the use of brain functions, which results in lower performance than expected. In contrast to the traditional psychometric model, increases in retest performance are expected to be more prominent than decreases, with the degree of improvement varying based on the psychological demands of the subtest.

The neuropsychological and educational literature bases have identified several potential mediators of retest improvement related to the psychological demands of the tasks, including task novelty and motor-speed requirements (Kaufman, 2003; Lezak et al., 2004). Table 14 presents the percentages of the total sample that showed positive, negative, and no test-retest scaled-score changes for all subtests along with the test-retest reliability coefficients derived from analysis of the data.

Table 14

Summary of Percentage of Cases within Score-Change Categories and Reliability Coefficients for Each WISC-IV Subtest and Selected Process Scores

Subtest	Negative Change	No Change	Positive Change	r_{x1x2}
Vocabulary	27.2	33.5	39.3	.92
Information	20.1	35.4	44.4	.89
Similarities	21.4	28.0	50.6	.86
Comprehension	29.6	26.1	44.5	.82
Word Reasoning	23.1	20.2	56.7	.82
Block Design	14.2	26.3	59.5	.82
Matrix Reasoning	21.3	30.1	48.5	.85
Picture Concepts	26.9	16.7	56.5	.76
Picture Completion	6.9	18.9	74.2	.84
Digit Span	24.8	28.2	46.8	.83
Digit Span Forward	25.4	30.0	44.5	.76
Digit Span Backward	29.1	31.2	39.6	.74
Letter Number	30.2	27.7	42.2	.83
Arithmetic	25.8	26.7	47.4	.79
Coding	15.6	16.9	67.6	.84
Symbol Search	19.7	18.5	61.7	.80
Cancellation	21.4	14.5	64.1	.79

For all subtests, the largest percentage of examinees demonstrated scaled-score increases of 1 or more points. Consistent with the literature base, subtests in the PRI and PSI indexes showed the greatest percentages of positive change. In general, the PRI and PSI subtests are novel and likely to elicit the development of new problem-solving strategies. It may be the use of these problem-solving strategies along with increases in motor and cognitive processing speed that account for the greater rates of retest improvement demonstrated for these subtests.

Of course, as previously indicated, not all subtests within a domain measure the same cognitive capacity. The Picture Completion Subtest, within the Perceptual Reasoning domain, demonstrated the highest rate of retest improvement among all subtests. Picture Completion requires the examinee to scan pictures and identify the missing essential component, but examinees are allowed only 20 seconds to scan each item. For this task, it is likely that robust retest gains are the result of the doubling of time of exposure to the pictures, which allows for more efficient scanning on the second administration and greater likelihood of identification of missing elements undetected during the first administration.

To understand the retest score patterns for subtests in the Verbal and Working Memory domains, it is necessary to examine the primary and secondary cognitive capacities measured by these subtests. As indicated in Table 5 the Verbal subtests, specifically Information and Vocabulary, yielded the highest percentages of no test-retest score change. These results are likely related to the memory demands of these tasks. Required for successful performance on both subtests is the ability to retrieve information from long-term storage. Long-term retrieval of verbal information is not a cognitive capacity likely to show extreme variation within a short period of time, such as a 2-4 week retest interval, because these subtests rely on accessing crystallized knowledge (i.e., knowledge acquired via formal and informal learning experiences). We would not expect such funds of stored information to be dramatically altered during the test-retest period. Minor variations in cognitive efficiency of retrieval, however, are much more common. As a result, scores on a second testing are more likely to show a larger number of minor fluctuations from the performance demonstrated at first testing.

As indicated in Table 10, the percentages of score regression and no change were particularly high for the Working Memory subtests. One possible explanation, at least for the Digit Span and Letter-Number Sequencing subtests, is that the stimuli to be recalled are decontextualized information units presented in random sequences that are not as easily handled during initial registration or as easily manipulated in working memory as information that is contextual and presented in a coherent, highly meaningful sequence, such as an arithmetic word problem or clues to the meaning of a word. The emphasis on random presentation of decontextualized information increases the potential for fluctuations in cognitive efficiency as well as the influence of random-error factors. Under these more random conditions of presentation and processing, score differences would be expected to be much more randomly distributed around the score earned during first testing.

As discussed in the review of the literature, minor variations in retest performance are expected for all subtests on the basis of variation in the use of executive function capacities (McCloskey, Perkins, & VanDivner, 2009; McCloskey, 2009a; McCloskey, 2009b; Denkla, 2007). For all subtests, efficient engagement of executive functions is necessary for consistent performance or improvements in performance across multiple administrations of a task. Conversely, variations in the use of executive function capacities that direct sustained attention, focused effort, and/or motivation can result in small or large variations in performance across multiple administrations of the same task.

Examination of the change category percentages along with the correlation coefficients provided in Table 14 highlights the differences between the clinical utility of the neuropsychologically based model and the traditional psychometric model. Providing an average of variability of scores from time 1 to time 2 in the form of a single reliability coefficient obscures the score variation patterns that are present in the data, thereby reducing clinical utility. The lack of a meaningful relationship between the score-change patterns of each subtest and the reliability coefficient value is very disconcerting as well. As a case in point, consider the fact that Vocabulary, a subtest considered to be one of the most reliable based on an obtained test-retest coefficient of .92, demonstrates a much more diffuse pattern of score changes (27.2% negative change, 33.5% no change, 39.3% positive change) than that of Cancellation, a subtest considered to be one of the least reliable based on an actual stability coefficient of .79. However, for Cancellation, there is a much less diffuse pattern of score changes (14.5% negative change, 21.4% no change, 64.1 percent positive change). Although the mathematical accuracy of the traditional psychometric method cannot be denied, the utility of the information conveyed by the reliability coefficients for these two subtests is difficult to comprehend, even for test users who have a good grasp of mathematics and test theory. Suggesting that the Vocabulary subtest is more reliable because a higher percentage of cases (33.5%) showed no change from time 1 to time 2 than is the case for Cancellation (14.5%) seems to miss the very important fact that the outcome of testing at time 2 is much more predictable for Cancellation than for Vocabulary in that a majority of cases (64.1%) performed better on the second testing and only a small minority (14.5%) performed less effectively. In the case of the Vocabulary subtest, the chances of increasing, decreasing, or staying the same on the second testing are nearly identical, making predictions much less accurate than in the case of the Cancellation subtest.

Showing the results of test-retest studies in the change category format used in Table 14 along with the reliability coefficients would be one way of increasing the clinical utility of the information offered about test-retest studies as discussed in the previous example. Other formats that reflect the range of the variability of the score changes, however, might be even more effective. In the case example just mentioned, knowing that the large majority of the variability of change for the Vocabulary subtest is contained within the range of -1 to +1 scaled-score points enables the clinician to appreciate the relatively stable nature of the Vocabulary subtest. It would also allow clinicians to understand that while Vocabulary scores may fluctuate somewhat unpredictably, the degree of fluctuation is negligible in terms of statistical or clinical significance. Knowing that the range of positive gains for the Cancellation subtest can vary greatly is equally important, but such knowledge should not diminish a test user's confidence in the Cancellation subtest as a reliable measure of processing speed. Rather, the data on range of score changes enable the clinician to anticipate the likely changes in a score profile if the test were re-administered within 2-4 weeks. The degree of consistency of the observed patterns of changes with the neuropsychologically based model predictions of variability of performance demonstrates the enhanced utility of such a model over the traditional psychometric model.

Limitations and Suggestions for Future Research

The present study represents an initial attempt at reconceptualizing the concept of reliability and the manner in which intelligence-test subtest reliability is presented. As such, future research is needed to refine this study's methodology and to extend its results.

A major limitation that may affect the generalizability of the findings was the test-retest interval, which averaged 32 days. With this short interval, it is difficult to establish the levels of retest performance increases that might be expected in actual clinical situations, where the test-retest interval is likely to be longer. Future research could address this limitation by evaluating changes in subtest performance at various time intervals. Of course, doing so will require consideration of other factors that may account for performance changes, such as maturation and other historical events.

Another limitation of this study was its failure to include larger samples of special populations who may demonstrate patterns of test-retest score changes different from those demonstrated by the standardization sample. The WISC-IV standardization sample excluded several subgroups, including those with limited English proficiency and those with a history of physical impairment that might depress performance, such as stroke, epilepsy, brain tumor, traumatic brain injury, history of brain surgery, encephalitis, and meningitis. Future research using the performance consistency model with these populations has the potential to further our understanding of the effects of different types of brain injuries and impairments on retest performance. Also, in cases where intelligence-test subtests are used as baseline measures and/or to measure progress or deterioration, the model presented here may provide a method for predicting rate of recovery based on subtest improvement.

Also necessary is additional research applying the performance consistency model to samples of learning-disabled (LD) children. Despite recent interest in alternative methods for learning-disability identification, such as the *Response to Intervention Model* (Gresham et al., 2005), the use of intellectual assessment instruments remains a common practice when evaluating children suspected of having learning disabilities. Furthermore, research by Fiorello et al., (2007) indicates that children with learning disabilities are more likely to show variable versus flat subtest profiles.

These findings suggest that future research is needed to identify the patterns of retest performance associated with LD subtypes. Such research has the potential of providing additional support for the idiographic approach to intelligence-test interpretation. It may also provide valuable information for practitioners who include cognitive assessment as part of the re-evaluation process or when testing procedures are unintentionally duplicated.

Finally, the present study demonstrated that a neuropsychologically based performance model is more effective at representing WISC-IV subtest test-retest reliability than are traditional psychometric procedures. In doing so, it provided both the rationale and the methodology for future studies to evaluate new methods of predicting retest performance based on knowledge of the psychological demands of the task.

References

Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

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

Bever, T.G., & Chiarello, R.J. (1974). Cerebral dominance in musicians and nonmusicians. *Science, 185,* 537-590.

Buckholdt, J.A. (2001). A short history of g: Psychometrics' most enduring and controversial construct. *Learning and Individual Differences, 13,* 101-114.

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

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.

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

Denckla, M.B. (2007). Executive function: Binding together the definitions of Attention Deficit/Hyperactivity Disorder and learning disabilities. In L. Meltzer (Ed.), *Executive function in education: From theory to practice*. New York: Guilford Press.

Dudek, F.J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335-337.

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

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

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

Flanagan, D.P., & Kaufman, A.S. (2009). Introduction and overview. In D.P. Flanagan & A.S. Kaufman (Eds.), *Essentials of WISC-IV assessment* (pp. 1-52). Hoboken, NJ: Wiley.

Flanagan, D.P., Ortiz, S.O., Alfonso, V.C., & Mascolo, J.T. (2002). *The achievement test desk reference (ATDR): Comprehensive assessment and learning disability.* Boston, MA: Allyn & Bacon.

Floyd, R.G., Evans, J.J., & McGrew, K.S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across school-age years. *Psychology in the Schools*, *40,* 155-171.

Glutting, J.J., McDermott, P.A., & Stanley, J.C. (1987). Resolving differences among methods of establishing confidence limits for test scores. *Educational and Psychological Measurement, 47,* 607-614.

Gold, J.M., Berman, K.F., Randolf, C., Goldberg, E., & Weinberger, D.R. (1996). PET validation of a novel prefrontal task: delayed response alteration. *Neuropsychology, 10,* 3-10.

Goldberg, E. (2001). *The executive brain*. New York, NY: Oxford University Press.

Goldberg, E. (2009). *The new executive brain: Frontal lobes in a complex world*. New York: Oxford University Press.

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

Goldberg, E., Harner, R., Lovell, M., Podell, K., & Riggio, S. (1994). Cognitive bias, functional cortical geometry, and the frontal lobes: Laterality, sex, and handedness. *Journal of Cognitive Neuroscience, 6*, 276-296.

Gottfredson, L.S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24,* 79-132.

Gottfredson, L.S. (1998). The general intelligence factor. *Scientific American Presents,* 24-29.

Gravetter, F.J. & Wallnau, L.B. (2007). Statistics for the behavioral sciences (7th ed.). Belmont, CA: Thomson-Wadsworth.

Greshman, F.M., Reschly, D.J., Tilly, W.D., Fletcher, J., Burns, M.K., Christ, T., et al. (2005). Comprehensive evaluation of learning disabilities: A response to intervention perspective. *School Psychologist,* 59(1), 26-30.

Guilford, J.P. (1954). *Psychometric methods (2nd ed.).* New York, NY: McGraw-Hill.

Hain, L. A., Hale, J. B., & Kendorski, J. G. (2009). The enigmatic population of specific learning disabilities: Comorbidty of psychopathology in cognitive and academic subtypes. In S.G. Feifer & G. Rattan (Eds.), Emotional disorders: A neuropsychological, psychopharmacological, and educational perspective. Middleton, MD: School Neuropsych Press.

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

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

Hale, J.B., Fiorello, C.A., Kavanagh, J.A., Hoeppner, J.B., & 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.

Heekeren, H.R., Marrett, S., Bandettini, P.A., & Ungerleider, L.G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature, 431,* 859-862.

Henson, R., Shallice, T., & Dolan, R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. *Science, 287,* 1269-1272.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B.K. Bryant (Eds.)*, Clinical neuropsychology and brain functions: Research, measurement, and practice* (pp. 125-167). Washington, DC: American Psychological Association.

Kaplan, E., Fein, D., Morris, R., Kramer, J.H., & Delis, D.C. (1999). The WISC-III as a processing instrument. San Antonio, TX: The Psychological Corporation.

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

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.

Kaye, D.B., & Baron, M.B. (1987). Long-term stability of intelligence and achievement scores in specific-learning-disabilities samples. *Journal of Psychoeducational Assessment, 3,* 257-266.

Kranzler, J.H. (2001). Commentary on “is g a viable construct for school psychology?” *Learning and Individual Differences, 13,* 189-195.

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

Lezak, M.D. (1988). IQ:RIP. *Journal of Clinical and Experimental Neuropsychology, 10,* 351-361.

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

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

Linacre, J.M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

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

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

Magnusson, D. (1967). *Test Theory*. Reading, MA: Addison-Wesley.

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.

Martin, A., Wiggs, C.L., & Weisberg, J. (1997). Modulation of human medial temporal lobe activity by form, meaning, and experience. *Hippocampus, 7,* 587-593.

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: The Psychological Corporation.

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

McCloskey, G. (2009a). The WISC-IV integrated. In D.P. Flanagan & A.S. Kaufman (Eds.), *Essentials of WISC-IV Assessment*. New York: John Wiley & Sons.

McCloskey, G. (2009b). 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*. Hoboken, NJ: Wiley.

McCloskey, G., & Maerlender, A. (2005). The WISC-IV integrated. In A. Prifitera, D.H. Saklofske, & L.G. Weiss (Eds.), *WISC-IV clinical use and interpretation: Scientist-practitioner perspectives* (pp. 101-149). Burlington, MA: Elsevier.

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

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

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

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometrics Monographs, 15*.

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

Miller, D.C. (2007). *Essentials of school neuropsychological assessment*. New York, NY: Wiley.

Miller, D.C., & Hale, J.B. (2008). Neuropsychological applications of the *WISC-IV* and *WISC-IV Integrated*. In A. Prifitera, D. Saklofske, & L. Weiss (Eds.), *WISC-IV* clinical use and interpretation: Scientist-practitioner perspectives (2nd ed.). New York, NY: Elsevier.

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

Prifitera, A., Weis, L.G., & Saklofske, D.H. (1998). The WISC-III in context. In A. Prifitera & D.H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist -practitioner perspectives* (pp. 1-38). San Diego, CA: Academic Press.

Psychological Corporation, The. (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). *Cerebral Cortex, 4,* 26.

Rasch, G. & Lord, F.M. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

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* (4th ed.). La Mesa, CA: Jerome M. Sattler.

Shadmehr, R., & Holcomb, H.H. (1997). Neural correlates of motor memory consolidation. *Science, 276,* 1272-1275.

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

Silver, N.C., & Dunlap, W.P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology, 72,* 146-148.

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

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.

Stanley, J.C. (1971). Reliability. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 356-442). Washington, DC: American Council on Education.

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

Suen, H.K., & Lei, P. (2007). Classical versus generalizability theory of measurement. *Educational Measurement, 4,* 1-13.

Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*, 263-267.

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

Traub, R.E., & Rowley, G.L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement*, *4*(4), 517-545.

Tulving, E., Markowitsch, H.J., Craik, F.E., Hiabib, R., & Houle, S. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cerebral Cortex*, *6,* 71-79.

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

Wechsler, D. (1955). *WAIS manual*. New York, NY: The Psychological Corporation.

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

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

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

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

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-Third Edition (WISC-III).* San Antonio, TX: The 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). *WechslerIntelligence Scale for Children-Fourth Edition (WISC-IV).* San Antonio, TX: Harcourt Assessment, Inc.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV)*. San Antonio, TX: Pearson.

Weiss, L., Saklofske, D., Prifitera, A., & Holdnack, J. (2006). *WISC-IV advanced clinical interpretation*. San Diego, CA: Academic Press.

Wright, B.D., & Masters, G.N. (1982). Rating scale analysis. Chicago, IL: MESA Press.