A RELIABLE APPROACH TO PSYCHOLOGICAL ASSESSMENT USING COGNITIVE TESTING BATTERIES

University of Southern Queensland

A Dissertation submitted by

Tammie Olm-Madden B. Sc. (Hons)

For the award of

Doctor of Philosophy

2008

ABSTRACT

Cognitive tests are rarely used in isolation. Instead the collection of tests into batteries is common place in clinical assessment. Clinical batteries range from fixed collections of tests administered unchanged to each client, to batteries flexibly constructed according to a process of hypothesis testing which varies between clients. Reviews of clinical practice indicate that clinicians predominantly employ a semi-flexibly constructed battery, comprised of a core group of measures with the addition of others drawn as needed from an available pool. While this accommodates for clinical concerns, the psychometric characteristics of such a battery tend to be unevaluated and clinicians draw inferences without reference to the resultant associated measurement error. This has been duly noted in the research literature which increasingly cites the need for psychometric evaluation at the battery level. The current investigation was undertaken to address this difficulty of clinical practice and aimed to develop a psychometrically and practically driven actuarial model with which practicing clinicians could structure and analyse cognitive batteries with due reference to reliability, validity and clinical utility. To this end, a review of psychometric literature was undertaken to determine theoretical guidelines for the control and measurement of error at the individual test and battery level. Reviews indicated that to successfully accommodate for the impact of random measurement error, clinicians must apply reliability theory to evaluation of the error associated with domain-based combinations of tests. Additionally, to ensure the validity of test-based inferences and avoid errors in decision-making, clinicians must apply empirically validated structures of cognitive function to guide test selection and combination. Given the pressing necessity of battery flexibility, it was concluded that clinicians could best accommodate psychometric and clinical factors

by the use of flexibly constructed composite scores. A reliable approach to psychological testing (RAPT) was proposed which applied psychometric theory and clinical factors to the development of a robust battery structure. The RAPT method focussed on the use of composite scores of domain-specific tests, grouped according to empirically validated domains and moderated by direct estimation of composite reliability. The RAPT was developed with the aim of facilitating the application of psychometric, actuarial methodology to a flexible collection of cognitive tests. In explicating the RAPT model, fifteen primary algorithms were derived from the psychometric literature and outlined according to 3 stages of battery usage: test selection; test analysis; and, test interpretation. The utility of the RAPT was examined in terms of its capacity to improve psychometric robustness within a flexible battery. Specifically, using simulated demonstrations, RAPT was demonstrated to provide a means of formalising empirically validated structure within a battery of tests, of controlling and improving the reliability of domain-based composite scores, of reducing the impact of artifactual errors on domain-based inferences and of applying actuarial methods typically associated with fixed batteries to a flexible collection of measures. Following this, RAPT was demonstrated to replicate existing psychometrically valid and stable interpretive structures. Specifically, RAPT algorithms were used to re-create the normative information provided for the Wide Range Achievement Test, Fourth Edition (WRAT-4) Reading Composite. Norms calculated using RAPT were compared with those provided in the WRAT-4 interpretive manual with minimal differences found. RAPT algorithms were then used to re-create normative and ipsative tables, summary scores intercorrelations, and reliability coefficients for the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) summary scores: Full Scale IQ (FSIQ); Verbal IQ

(VIQ); Performance IQ (PIQ); Verbal Comprehension Index (VCI); Perceptual Organisation Index (POI); Processing Speed Index (PSI); and, Working Memory Index (WMI). Again, RAPT was demonstrated to successfully replicate these data. Finally, the clinical utility of RAPT was demonstrated with the illustration of case examples which outlined the capacity of RAPT to apply psychometrically sound methodology to the tasks of modifying existing composites, modifying existing battery structure and developing battery structure and analyses according to a factor analysis investigation of an Australian normative sample (n=1045). These investigations provided comprehensive evidence of the capacity of the RAPT to enable the direct application of psychometric theory to semi-flexible battery construction in a manner suitable for use in current clinical practice.

CERTIFICATE OF DISSERTATION

I certify that the ideas, experimental work, results, analyses and conclusions reported in this dissertation are currently my own effort, except where otherwise acknowledged. I also certify that the work is original and has not been previously submitted for any other award except where otherwise acknowledged:

_____ _____

Signature of Candidate Date

ENDORSMENT

_____ _____

Signature of Supervisor Date

ACKNOWLEDGEMENTS

First to my David who has infinite love, patience, kindness and strength and of whose ultimate generosity I have most frequently availed me. You have given so freely, without ever stinting. There are not enough words to explain how completely you are my life and how good it is, my husband, “on this side”.

To Kristine, my mother, who in not striving for herself strove instead for us and in so doing provided an impetus which was necessary and priceless and which I have entirely relied on.

To Daddy who loves me as I am and who has provided a foundation of strength that he himself would not believe or credit. This has sustained me many times.

To Joseph, Paxie, Aleacia and David, my sibling walkers on this uphill pathway, in fellowship. With all my honest love.

To Graeme, who is occasionally infuriating but who is also unfailingly kind, whose words I have trouble remembering are not my own and without whom I would never have learned to re-build a city according to a plan of my own.
In gratitude for my apprenticeship.

To the Madden family for welcoming me as one of their own and for continually inspiring me with their own achievements. With much love.

To my friends, Heather, Hong Eng, Yong Wah and Mark, who have been my comrades and who are my philosophers.

And, to Nana, for her history, for her humour and her quiet encouragement.

In hope that these few words will best carry the depth of my thanks
I am most sincerely grateful to you and all the others.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FORMULAE

CHAPTER ONE

INTRODUCTION

1.1 Introduction

It would be the rare clinician who considers a cognitive assessment to be complete after administering only one test. To the contrary, assessment typically involves the use of multiple measures with the aim of eliciting a wide range of relevant behaviours. Testing batteries provide substantially more information regarding the cognitive functioning of an individual than the use of any single test. However, the difficulties of developing batteries that are both psychometrically robust and clinically appropriate generate substantial challenges to practicing clinicians.

Test batteries may be entirely "flexibly" constructed to include unique collections of cognitive measures in a process of hypothesis testing. Potentially this can result in batteries that are as diverse as the clients and conditions they are constructed to assess. Alternately, they may be "fixed" or unchanging in their make-up. With a fixed battery the same combination of measures is administered to all clients regardless of their individual circumstances or needs. Most commonly, however, a battery of cognitive tests is comprised of measures chosen flexibly from a limited pool ("semi-flexible"), with clinicians tending to favour a common group of measures with which they are familiar and to which they have ready access. Regardless of the method of construction, a battery of cognitive tests must be subjected to considerations of accuracy, stability and meaningfulness if it is to be used to draw inferences about the cognitive functioning of an individual.

A wealth of psychometric literature provides methodologies for the quantification and control of the measurement error associated with the use of

psychological tests. Be that as it may, a gulf exists between psychometric theory and its application to actual clinical practice. The selection and use of psychological tests may be influenced more by those measures that psychologists were trained to use in their respective graduate programs, the commercial availability and cost of measures, and the policies and funding of institutions that employ assessment psychologists. Nowhere is this gulf more evident than when assessment requires the use of multiple tests where the potential sources of error are manifold. While psychometric literature frequently pertains to the control and measurement of error at the level of the individual test, very little test theory relates to the control and measurement of errors which occur when analysing and interpreting the results of a test battery as a whole. This is despite the fact that combining tests into a battery in no way reduces the potential for measurement error associated with both the individual measures and the battery as a whole. Scrutiny of these issues is the purpose of the current thesis which was conducted with the aim of proposing methodology for the control and measurement of error when semi-flexible batteries of cognitive tests are used.

1.2 Current Status of Battery Usage

Clinicians predominantly use multi-measure batteries when describing the cognitive functioning of adults in forensic, educational and clinical settings (Rojas & Bennet, 1995; Sweet, Moberg & Suchy, 2000). Lees-Haley, Smith, Williams and Dunn (1995) in a review of a hundred neuropsychological reports written for forensic referrals found that on average clinicians administered over eleven tests (M = 11.7) with as many as thirty-two instruments per battery. Similarly, Camara, Nathan and Puente (2000) reported that clinical psychologists used an average of thirteen (M = 13.4) individual tests in an assessment, while neuropsychologists used an average of

seventeen and as many as thirty different tests per assessment (M = 17.6). Most recently, clinicians participating in a survey by Rabin, Barr and Burton (2005) cited an average of twelve individual instruments per battery. Such reviews indicate that use of a multi-measure battery is common, however, there is little consensus regarding how the different measures contribute to the role of the battery as a whole, and less regarding the correct methods of constructing, analysing and evaluating such a battery (Bauer, 2000; Lezak, 1995; Russell, 2000).

In the clinical literature, methods of battery construction differ widely in terms of structure, test selection, and the means by which the resulting test scores are combined, analysed and interpreted (Bauer, 2000; Lezak, Howieson & Loring, 2004; Milberg, Hebben & Kaplan, 1996; Reitan & Wolfson, 1996; Russell, Russell & Hill, 2005). Battery construction techniques in clinical practice fall on a continuum between use of an unchanging or "fixed" battery, applied in every instance without modification, and use of an entirely "flexible" collection of measures, modified specifically for each individual test taker in order to pursue specific clinical hypotheses which arise in the course of investigation (Vanderploeg, 2000). Battery construction techniques differ fundamentally in practical terms (i.e., time and method of test selection), psychometric application (i.e., evaluation of psychometric reliability and validity at the battery level), theoretical focus (i.e., application of quantitative analysis versus application of a process of hypothesis testing) and historical development (i.e., formation based on fundamentally differing theories of cognitive functioning) (Bauer, 2000; Goldstein, 1997; Kane, 1991; Lezak, 1995; McKenna & Warrington, 1996; Milberg, Hebben & Kaplan, 1996; Reitan & Wolfson, 1996; Tranel, 1996). Underpinning debates about the appropriate mode of battery usage, however, are fundamental questions regarding the aims of assessment,

the theory underlying elicitation of specific cognitive behaviours and the intended application of test results (Kolb & Whishaw, 1996).

While exponents disagree regarding the appropriate construction and usage of cognitive testing batteries and the degree to which psychometric theory may be applied at the battery level, their often vitriolic discussions consistently fail to suggest clear and definitive recommendations for the practicing clinician. Despite the rhetoric, to date, no battery construction approach has been empirically demonstrated to produce more reliable, valid or clinically meaningful conclusions than any other (Bauer, 2000; Goldstein, 1997; Kane, 1991; Milberg, Hebben & Kaplan, 1996; Mitrushina, Boone, & D'Elia, 1999; Russell, Russell & Hill, 2005). The apparent irrelevance of theoretical battery research in practical terms is further emphasised by consideration of how batteries are used clinically. In reality, most clinicians undertake the complex and multi-stage task of establishing the depth and breadth of cognitive assessment by choosing from the available cognitive tests with widely differing degrees of reference to psychometric, cognitive, or neuropsychological theory. Reviews of clinical practice indicate that while clinicians may endorse a "fixed" orientation, few clinicians use a comprehensive fixed battery without modification. On the other hand, the marked similarities between the tests which are "commonly used" by clinicians indicate that few, if any, employ a purely flexible approach.

In a survey of 614 practicing neuropsychologists in North American (Seretny, Dean, Gray & Hartlage, 1986), ninety-five percent (95%) of the sample reported use of the "age appropriate Wechsler Scale", almost two thirds (63%) used "some portion" of the Halstead Reitan Neuropsychological Battery (HRNB; Reitan & Wolfson, 1985) and a third (35%) used "some portion" of the Luria Nebraska

Neuropsychological Battery (LNNB; Golden, Purisch & Hammeke, 1988). Rather than using a purely fixed or purely flexible approach, clinicians tended to supplement fixed battery subtests with additional measures of reading, visuospatial processing, memory and executive functioning.

A similar trend of semi-flexible battery construction was indicated in a survey of 449 clinicians conducted by Guilmette, Faust, Hart and Arkes (1990). In this study, the "Wechsler Intelligence Scales" and "Wechsler Memory Scales" were reported by clinicians as the most frequently used measures, while fewer clinicians used the full HRNB (27%) or the LNNB (18%). Again, however, clinicians appeared to modify the core battery (i.e., the WAIS or WAIS-R and WMS or WMS-R) with additional cognitive tests.

Butler, Retzlaff and Vanderploeg (1991), in a review of 250 practicing neuropsychologists, reported that a majority of clinicians used the Wechsler Adult Intelligence Scale (i.e., WAIS) or WAIS-R (86%), while a smaller number used the HRNB (29%) or the LNNB (8%). As with previous studies, clinicians in this study typically administered tests of memory, language, visuospatial and executive functioning in addition to the core batteries.

The prevalence of semi-flexible battery construction was indicated by Lees-Haley, Smith, Williams and Dunn's (1995) review of 100 forensic neuropsychological reports in which battery construction varied between all reports in the study. No two reports in this review contained exactly the same battery, convincingly challenging claims of "fixed" battery usage by the participating clinicians.

Use of fixed batteries is even less common amongst Australian psychologists. Sullivan and Bowden (1997) in a comprehensive review of Australian clinicians

reported that less than ten percent (10%) of clinicians sampled used the HRNB and none used the LNNB. In fact, a majority of clinicians in this survey used a core battery (WAIS or WAIS-R, and WMS-R) with additional tests flexibly drawn from a pool of some eighteen other commonly used measures.

This trend has been further demonstrated through the clinical practice reviews of Sweet, Moberg and colleagues (Sweet, Nelson & Moberg, 2006; Sweet, Moberg & Suchy, 2000; Sweet, Moberg & Westergaard, 1996; Sweet & Moberg, 1990) who conducted four consecutive surveys of practicing clinical neuropsychologists over a sixteen year period and reported universal increases in the use of a semi-flexible battery along with almost consistent declines in use of purely fixed or purely flexible test groups (see Table 1.1). It is notable that given the date of the most recent of these surveys, this longitudinal review of clinical practice adds highly relevant information on current battery usage.

Table 1.1

Summary of Practice Reviews

		Battery Type		
	N	**Semi-flexible**	**Fixed**	**Flexible**
Sweet & Moberg, 1990	184	54%	18%	29%
Sweet, Moberg & Westergaard, 1996	279	60%	14%	25%
Sweet, Moberg & Suchy, 2000	420	70%	15%	16%
Sweet, Nelson & Moberg, 2006	1078	76%	7%	18%

Despite the wealth of research, neither a fixed nor a flexible approach produces consistently superior clinical conclusions, in terms of reliability, validity,

clinical meaningfulness and adaptability. In fact, debate between adherents is conducted largely on ideological rather than experimental grounds (Bauer, 2000; Franzen, 2000; Goldstein, 1997; Kane, 1991; Lezak, Howieson & Loring, 2004; Mitrushina, Boone & D'Elia, 1999; Russell, Russell & Hill, 2005). Commonly accepted outcomes such as diagnosis relative to external criteria, successful rehabilitation planning and accurate prognosis or behavioural prediction, convincingly fail to indicate the superiority of either approach and when such arguments are made they are largely anecdotal and counterbalanced (Goldstein, 1997). The irrelevance of the fixed versus flexible debate in practical terms is emphatically underscored by reviews which indicate that clinicians do not typically use either method in its purest form (Sweet & Moberg, 1990; Sweet, Moberg & Suchy, 2000; Sweet, Moberg & Westergaard, 1996; Sweet, Nelson & Moberg, 2006).

The task of constructing a cognitive battery is complex and perhaps compounded by the lack of consensus regarding methodology. Clinicians must determine the scope of cognitive abilities to be examined, choose appropriate tests, and determine the depth of investigation with reference to time constraints, referral questions, individual requirements of the client, developments in test research and a myriad of other concerns (Bauer, 2000; Lezak, Howieson & Loring, 2004; Vanderploeg, 2000). Clinical reviews convincingly indicate that practicing clinicians attempt this delicate balance while using predominantly semi-flexible batteries. Psychometric literature suggests, however, that in doing so clinicians risk incorporating unquantified degrees of measurement error.

1.3 Measurement Error Using Individual Tests

All measurement contains error and failure to evaluate or control this error does not remove it from any set of scores (Feldt & Brennan, 1993). Because error is endemic to test scores, ultimately it must be considered for the quantification of human behaviour to have meaning (Lord & Novick, 1968).

While the use of multi-measure, semi-flexible batteries is the norm in clinical practice, lack of psychometric evaluation, insufficient consideration of clinical constraints and failure to apply systematic methods of structuring and analysing the battery, introduce error into the quantification of cognitive behaviours. This failure to effectively control for or adequately measure error occurs at the level of both the individual test and the battery and potentially compromises the efficacy of cognitive assessment as it is currently conducted.

The process of assigning interpretable numerical values to cognitive behaviours using test scores requires two fundamental evaluations. The first of these is determination of the degree to which the underlying construct is accurately quantified by the test score: that is, the "meaning" of the test score in behavioural terms. The second is determination of the dependability of the test score itself: that is the degree to which the score itself may be trusted to be robust to the impact of confounding factors, and hence to be replicable. These considerations are fundamentally based on the degree to which measurement incorporates error.

Most cognitive tests are intended to measure stable traits or ability, such as language ability, memory, attention, visuospatial ability, or executive functioning (Lezak, Howieson & Loring, 2004; Mitrushina, Boone & D'Elia, 2005; Spreen & Strauss, 1998). In using test scores to quantify such behaviours, the clinician trusts that the scores he or she obtains are accurate measures of the construct of interest.

Variation within test scores cannot, however, be solely attributable to the relevant construct. Thorndike (1951), for example, categorised potential sources of variation within an observed score into variation due to permanent characteristics of the individual (i.e., consistent level of ability in the construct of interest or general ability in test taking), variation due to temporary characteristics of the individual (i.e., health, fatigue, motivation, emotionality or practice effects), or variation due to other factors which are likely to be transient, such as luck, guessing or distraction. Variability due to irrelevant factors obfuscates consistent measurement and accurate understanding of the underlying trait: in other words introduces error into the measurement.

This directly impacts upon clinical inferences which are the primary aim of measurement. Cognitive test scores provide estimates of functioning on constructs from which true levels of functioning are inferred. If the observation is confounded by factors other than the construct of interest, the accuracy of this inference is undermined. The process of quantification inevitably incorporates such confounds and most test scores contain some element of "truth" and some element of "error". This is unavoidable. However, if the error component of any given observed score is too great the process of drawing inferences regarding true levels of functioning from the score is fundamentally flawed (Nunnally, 1978). For this reason a comprehensive understanding of measurement error is of vital importance to clinicians who use test scores to draw just such inferences.

Feldt and Brennan (1993) defined measurement error as variations or "inconsistencies" within a test score not caused by the construct of interest. These theorists summarised four primary sources of error variation relating to the test taker, the test setting, the clinician and the measurement instrument. These are outlined in

Table 1.2 below which summarises error sources defined initially by Thorndike (1951), elaborated upon by Stanley (1971) and subsequently endorsed by Feldt and Brennan (1993).

Table 1. 2

Sources of Error in Measurement

Source of Error	**Cause**	**Controlled/Measured**
Random variation within each individual	Fluctuations in health, motivation, mental efficiency, concentration, care and luck.	Must be considered in interpretation and controlled for by appropriateness of norms.
Situational Factors	Fluctuations in working environment, psychological reaction to test taking (i.e., anxiety), physical ability to complete testing.	Must be controlled for by test selection and standardised administration.
Examiner Factors	Deviations from standardised administration, impact of examiner personality on the individual, impact of observation on test performance.	Must be controlled for by standardised administration and objective scoring. May be measured using reliability (i.e., inter-rater).
Instrument Factors	Instability intrinsic to the test.	Must be measured using reliability (i.e., internal consistency, test-retest).

Contemporary psychometric texts further simplify the sources of test score variation into variation due to stable characteristics of the individual or test setting which relate to the construct of interest and contribute to test score consistency, and variation due to transient characteristics of the individual or test setting which contribute to test score inconsistency and which do not pertain to the construct of interest (Murphy & Davidshofer, 2005). The distinction between error sources which exert a systematic influence on measurement and those which exert a non-systematic, or random, influence is essential to the clinical application of psychometric theory as it determines which sources of error may be controlled and which must be, at least, measured.

As indicated in Table 1.2, error associated with the individual and the setting may be controlled by clinician behaviours such as the use of standardised test administration, building of adequate rapport between client and clinician, and minimising distractions within the testing environment. Error introduced by the examiner may also be controlled or may be evaluated in part by estimation of the typical stability of judgements on a test across examiners (Anastasi & Urbina, 1997). Error, which is inherent to the test itself, may be measured using psychometric estimates of the test reliability (Feldt & Brennan, 1993). Regardless of how the sources of error are summarised, however, to the extent that error cannot be controlled it must be measured if the test user wishes to "portray nature in its ultimate lawfulness" (Nunnally, 1978, p. 191).

The concept that all measurement includes error is a basic premise of classical test theory (CTT), the axioms of which are comprehensively derived in Gulliksen (1950), Lord (1959), Novick (1966), and Lord and Novick (1968). While other psychometric theories have been subsequently proposed, for example

generalizability theory (Brennan, 1983; Cronbach, Gleser, Nanda & Rajaratnum, 1972; Shavelson & Webb, 1991), and item response theory (Embretson & Reise, 2000), CTT provides many of the mathematical and conceptual means by which the impact of error on test scores is evaluated in modern instruments of cognitive assessment (Bechger, Maris, Verstralen & Beguin, 2003; Franzen, 2000). In part, this is because the fundamental assumptions of CTT are modest, and hence readily satisfied, and the theory is, therefore, widely applicable (Novick, 1966; Lord & Novick, 1968).

The usefulness of CTT to the test user is as a primary means of estimating the impact of measurement error on test scores. In CTT, error (i.e., "E") is conceptualised as the discrepancy between an observed score and the "true" score on an instrument for any given observation, thus:

$$E = X - T \qquad \text{Formula 1.1}$$

In formula 1.1, which is likely to be familiar to psychometricians and assessment clinicians alike, "X" indicates an observed test score and "T" indicates the true test score, which is conceptualised as the average score to be obtained by an individual from infinite repetitions of the test, assuming that all other factors remain constant (Novick, 1966; Lord & Novick, 1968).

If the scores derived from a particular measure are robust to the impact of error, measurement is consistent and replicable, that is to say reliable (Nunnally, 1978). Given the definition of true score above, it also follows that a test score is "meaningful" only to the extent that the true score actually represents the construct of interest. Consideration of both test score stability (i.e., reliability) and meaningfulness (i.e., validity) is the primary goal of psychometric theory (Franzen, 2000).

In practical terms, the estimation of measurement error through coefficients of reliability allows the clinician to establish the degree of confidence in a single score that is warranted by the estimated stability of that score across time, raters and forms. Accordingly, the impact of random error sources on measurement is estimated by reliability coefficients based on the CTT premise that the reliability of an instrument is a ratio of the variance associated with the true scores over the variance associated with the observed scores, which is expressed as follows:

$$r = \frac{\sigma^2_{True}}{\sigma^2_{Observed}} \qquad \text{Formula 1.2}$$

The assumption implicit in this conceptualisation is that the variance of observed scores taken from a group of individuals who actually differ only in terms of true scores, provides an expression of the impact of random error in measurement (Lord & Novick, 1968; Novick, 1966; Stanley, 1971; Streiner, 2003).

The reliability coefficient associated with use of a particular test, in a particular setting, with a specific group of individuals is used to derive an estimate of the average measurement error (i.e., standard error of estimate, measurement and prediction) and subsequently the interval around the observed score within which the true score is expected to fall with a particular level of certainty (i.e., the confidence interval). Thus clinical decision-making using cognitive test scores can be undertaken with direct reference to the estimated reliability of these scores.

At this point, it is necessary to note that the "true" score as defined in CTT differs from the component of "truth" in the observed score discussed in the preceding section (Lord & Novick, 1968; Streiner, 2003). This "truth" is defined as the individual's unique and intrinsic capacity on the construct of interest, with variation due to all other sources consigned to "error". As Lord and Novick (1968)

comment, this "Platonic" definition of true score is less applicable to CTT due to the "inexact constructs" upon which it is the business of the theory to comment. To the extent that the true score is a pure measure of the construct of interest, factors which are unrelated to the construct of interest provide the only occlusion of "truth". This may be expressed by formula 1.3, as follows:

$$X = CI + [SE + RE] \qquad \text{Formula 1.3}$$

Where "X" is an observed score, "CI" is the proportion of the score which relates specifically to the construct of interest, "SE" is the proportion of the observed score due to error factors which vary systematically with the CI (i.e., systematic error), and "RE" is the proportion of the observed score due to factors which are unrelated to the CI.

When measuring psychological constructs, however, the "true" score in CTT terms is rarely an unadulterated measure of the construct of interest. Nor, as an aside, is the construct of interest likely to be conveniently unidimensional. Instead the CTT conception of "true score" may incorporate the systematic influence of factors other than the construct of interest (i.e. see Lord, 1959; Novick, 1966; Lord & Novick, 1968; Judd, Smith, & Kidder, 1991), which be may expressed by a slight modification of formula 1.3, as follows:

$$X = [CI + SE] + RE \qquad \text{Formula 1.4}$$

As indicated in the formula, the aspect of the observed score, "X" which is unique from random error variance "RE", is in fact a function of both the construct of interest, "CI", and factors which vary systematically with this construct, "SE", and thus are difficult to isolate. In reality, psychological constructs are rarely purely defined and instead the "true score" is the result of both the relevant construct and similarly varying factors (i.e. see Streiner, 2003 for further discussion). The impact

of these factors on measurement is not considered under reliability and must be the basis of a second investigation, that of validity, which aims to determine the influence of systematic error factors by defining the true score. Streiner (2003) demonstrates this succinctly by incorporating these two components of "true score" into the CTT ratio defining reliability (formula 1.2) as follows:

$$r = \frac{\sigma^2_{CI} + \sigma^2_{SE}}{\sigma^2_{Observed}} \qquad \text{Formula 1.5}$$

In other words, reliability is a function of the ratio of the variance due to the construct of interest and systematic error, over the total variance of the observed score (i.e., where" "σ^2_{CI}" is the variance of the construct of interest, "σ^2_{SE}" is the variance of the systematic error component and "$\sigma^2_{Observed}$" is the total variance). On the other hand, validity is defined as:

$$v = \frac{\sigma^2_{CI}}{\sigma^2_{Observed}} \qquad \text{Formula 1.6}$$

That is, the ratio of the variance due only to the construct of interest, over the total variance of the observed score. While random error is a component of the denominator in both equations, systematic error exerts an influence only on validity. For example, in formula 1.5, systematic error is a component of both the numerator and denominator of the equation. The product of this ratio is a function of the random error which is incorporated into the total observed score variance, but not into the combination of true score and systematic error variance in the numerator. On the other hand, the ratio expressing validity (formula 1.6) again includes true score, systematic error and random error variance combined into observed score variance in the denominator, however, lists only variance associated with the true score in the numerator. The product of this ratio, therefore, is a function of both random and systematic error variance implicit to the observed score variance, but

removed from the true score variance in the numerator. CTT investigations of score reliability focus primarily on isolating random variances from those which exert systematic influence, regardless of whether these are related to systematic error or true score. This has arguably been achieved with some degree of mathematical precision. On the other hand, validity investigations focus on isolating the illusive "true" score: the aspect of the observed score which is free from both systematic and random error variance. That the latter investigation adds complexity is demonstrated by the decreased levels of mathematical precision in methodologies currently developed for its explication (Franzen, 2000).

As this discussion highlights, evaluation of the meaningfulness of individual scores is less succinctly demonstrated than evaluation of their stability and the test score is often taken to define the "construct of interest" exemplified by the infamously tautological comment that "intelligence is what is measured by intelligence tests" (Boring, 1923, p.36). In CTT, validity is often defined by the relationship between the scores of two tests: one which is commonly accepted as a measure of a specific construct; and the other which is undergoing validation. In this instance the two test scores are assumed to have the same true score but are differentially vulnerable to both random and systematic errors. Comparison aims to highlight the aspect of both scores attributable solely to true score. This fairly simplistic definition may be expanded to encompass definition of an underlying "construct" or "factor" of which the true score is a partial measure and which may be defined by various shared variance methods, frequently factor analysis (Delis, Jacobson, Bondi, Hamilton & Salmon, 2003; Thompson, 2004).

1.4 Measurement Error at the Level of the Battery

Psychometric theory, as outlined in the previous section, is by and large concerned with the control and measurement of error at the level of the individual test. Modern cognitive assessment typically requires a "complex, multifaceted" battery from which the clinician must draw a comprehensive description of the cognitive functioning of an individual (Vanderploeg, 2000). As discussed above, in practice most clinicians use multi-measure batteries to comprehensively sample a wide range of relevant cognitive abilities and to avoid errors to which single tests are particularly vulnerable (Cimino, 2000; Thorndike, 1951). Drawing inferences from a single test score incorporates error. It follows that drawing inferences from a collection of test scores is also vulnerable to error though perhaps of a different and more complex nature. In the task of interpreting a battery of test scores with often undefined relationships, clinicians must contend with unique and additional error sources which are arguably not considered in cognitive assessment to an appropriate degree.

While the impact of error on individual test scores is widely discussed (i.e. Anastasi & Urbina, 1997; Cronbach & Meehl, 1950; Dunnette & Borman, 1979; Foster & Cone, 1995; Franzen, 2000; Lord & Novick, 1968; Murphy & Davidshofer, 2005), evaluation of psychometric error at the level of the battery has not been accomplished to any great extent. Consequently, semi-flexible battery construction is often undertaken with little application of formal methods of evaluating and controlling for error. This has direct implications for the veracity of the resulting inferences (Garb & Schramke, 1996). In fact, current battery usage is vulnerable to errors due to failure to select and group tests systematically and according to a verified structure of cognitive functioning, failure to recognise and accurately

accommodate for clinical constraints, and inadequate evaluation of reliability, among other factors. In fact, compiling tests into a flexible battery without psychometric justification results in several errors including, inflated error rates, multicollinearity, weighting decision problems, unknown reliabilities, and variability across interpreters, among other problems (Ingraham & Aikken, 1996; Wedding & Faust, 1989).

Considerable research suggests that clinicians attempting to analyse the data from a multi-measure battery will be impeded by several human biases endemic to clinical judgements (Wedding & Faust, 1989; Williams, 1997). Such biases include hindsight bias resulting from the tendency to inflate estimations of accuracy in retrospect (Fischhoff, 1975; Mitchell & Kalb, 1981), and confirmatory bias caused by seeking data within the test battery to confirm a favoured hypothesis at the expense of conflicting data (Greenwald, Pratkanis, Leippe & Baumgardner, 1986; Mahoney, 1977). Over-reliance on salient or dramatic indicators of deficits (Wedding & Faust, 1989), under-utilization of base rates (Rosenfeld, Sands & Van Gorp, 2000), and failure to analyse co-variation, especially reliance on "illusory correlations" or unsubstantiated relationships between test data (Reitan & Wolfson, 2001) also limit interpretation of the typical clinical battery.

In addition, various human cognitive limitations threaten the validity of clinical judgements, including a limited capacity to process the many variables in an average data set (Dawes & Corrigan, 1974), over-reliance on memory (Arkes, 1981), overconfidence in judgements (Wedding & Faust, 1989), and errors in decision making strategies, such as the inability to eliminate preconceived notions (Chapman & Chapman, 1967). These errors of clinical judgement influence the battery at the structural, analytical and inferential levels (Williams, 1997) a fact which has been

increasingly considered by commentators on clinical assessment (i.e. Cimino, 2000; Crossen, 2000; Vanderploeg, 2000).

The semi-flexible battery has been recently criticised in the literature for its vulnerability to these errors. In a review of forensic assessment practice, Russell and Russell (2003) indicate the likely prevalence of clinical decision errors associated with use of a semi-flexible collection of tests. Errors cited by these authors included the misattribution of impairment based on a misinterpretation of normal variation between battery scores, use of insufficient tests and subsequent failure to elicit relevant behaviours, use of test scores to justify a pre-existing diagnosis, administration of tests until impairment was found, use of inappropriately easy or difficult tests, comparisons between tests with non-equivalent normative groups, failure to use norms corrected for age, education and gender, discrepancies between norms, lack of co-norming, and use of "un-validated" test batteries. The authors considered that given the lack of standardised, actuarial methods at the structural, analytic and interpretive levels of flexible and semi-flexible battery usage, errors in inference were a foregone conclusion (Russell & Russell, 2003). In light of such criticism Russell, Russell and Hill (2005) contended that the application of psychometric principles to the evaluation of the battery as a whole was necessary to ensure accurate and dependable inferences. These authors and other supporters of the fixed battery approach conclude that adequate control of measurement error is impossible with a flexible or semi-flexible collection of tests (i.e. Reed, 1999; Russell & Russell, 2003; Russell, Russell & Hill, 2005). In fact these and other exponents of fixed battery usage contend that the use of a standardised battery constitutes the most complete application of psychometric principles to the

evaluation of cognitive functioning at the battery level (Reitan & Wolfson, 1996; Russell & Russell, 2000; Russell, Russell & Hill, 2005).

This is clearly not the opinion of a majority of theoreticians (i.e. Bauer, 2000; Lezak, Howieson & Loring, 2004; Milberg, Hebben & Kaplan, 1996) or clinicians (i.e. Butler, Retzlaff & Vanderploeg, 1991; Camara, Nathan & Puente, 2000; Guilmette, Faust, Hart & Arkes, 1990; Lees-Haley, Smith, Williams & Dunn, 1995; Rabin, Barr & Burton, 2005; Seretny, Dean, Gray & Hartlage, 1986; Sullivan & Bowden, 1997; Sweet & Moburg, 1990; Sweet, Moburg & Suchy, 2000; Sweet, Moberg & Westergaard, 1996; Sweet, Nelson & Moberg, 2006) nor is it perhaps a justifiable position. Such criticisms of battery usage have, however, been useful in catalysing closer scrutiny of the error inherent to current cognitive assessment and in forwarding the solution of psychometric verification at the battery level.

Clinicians seem to consider that comprehensive fixed batteries fail to provide the answer to measurement error. The ubiquitous application of certain standardised collections of tests, such as the Wechsler intelligence and memory scales (Wechsler, 1939, 1981, 1997a, 1997b, 1997c), however, arguably demonstrates attempts to capitalise on the known and demonstrable strengths of a fixed collection of cognitive measures. Specifically, the fixed battery lends itself to several primary strategies which may be used to control measurement error. First, a fixed battery may be developed according to a specific theory of cognitive functioning, or for a specific purpose (i.e., to evaluate cognitive functioning or diagnose impairment). Such a process lends itself to empirical consideration of validity and the use of clearly defined underlying structure upon which tests are chosen and combined reduces the likelihood of several of the clinical judgement errors discussed previously. Secondly, the battery may be normed as a whole, eliminating error due to inequality

of norms and facilitating actuarial methods of analysis. On a related note, relationships between individual measures may be thoroughly evaluated, facilitating ipsative comparisons. Finally, a fixed and unchanging battery lends itself to thorough psychometric evaluation in terms of reliability and validity for various purposes. The battery may be validated, as a whole, according to its ability to achieve the purpose for which it was designed and according to adherence to the theory upon which it was based. Similarly, the reliability of the battery may be readily evaluated. Theoretically, the unchanging nature of the fixed battery facilitates psychometric strength in both normative and ipsative comparisons and alleviates many of the measurement errors endemic to the use of multiple measures.

Unfortunately, these potential strengths are of little use to the majority of clinicians who employ semi-flexible collections of cognitive measures, often for very appropriate and relevant clinical reasons (Cimino, 2000). Arguably, the available fixed batteries such as the HRNB (Reitan & Wolfson, 1985) and the LNNB (Golden, Purisch & Hammeke, 1988) provide neither the scope nor the flexibility to render them useful and widely used clinical measures. As discussed earlier, clinicians typically modify these batteries, fundamentally changing their psychometric characteristics. As Reitan and Wolfson (1985) state "The only authorized version of the HRNB is the one that duplicates the tests exactly as they were when the validation studies were done" (p.40): a form which clinicians rarely employ unchanged or in its entirety.

Perhaps the answer to measurement error does not lie in closer adherence to existing methods, but in the development of an alternative model of battery construction which integrates both psychometric and clinical requirements, and incorporates the strengths of both fixed and flexible construction. Somewhat

ironically, the multi-measure battery, if it is structured and analysed with due reference to the potential sources of error, in itself provides a solution to psychometric unreliability and invalidity, regardless of the flexibility of its construction. In order for the errors inherent to multi-measure batteries to be controlled, however, psychometric theory must be applied at the constructional, analytical and interpretive levels of battery usage. Consideration of an appropriate methodology to achieve this aim is the focus of the current research.

1.5 Summary of Thesis

"The conclusions reached via numerical argument must be conclusions that are wholly implied by the empirical data itself and not conclusions whose content depends on the numbers assigned" (Michell, 1986, p. 401). To this aim, the current research will investigate the impact of reliability, validity and clinical utility on the structure, analysis and interpretation of semi-flexible cognitive batteries. Investigations are based on the premise that meaningful, dependable and accurate inferences may only be drawn from battery scores when the sources of error are controlled and analysed using psychometric and clinical means. Regardless of the "statistical finesse and mathematical virtuosity" (Cattell, 1964, p.1) of psychometric theory, if it is not applied to actual clinical practice, the accuracy and meaningfulness of behavioural quantifications are in no way improved solely by theorising.

This thesis will be somewhat atypical in structure as its primary goal is to review relevant psychometric theory, and integrate findings into a methodology of battery construction and analysis that is readily accessible to clinicians and adapted for ready use in clinical practice. While much of the psychometric theory being considered is conventional, longstanding and generally incorporated in test

construction, there is little evidence that the lessons learned regarding measurement error are routinely employed in clinical practice. As discussed, assessment psychologists typically administer multi-measure batteries. However, interpretative methods and approaches that are informed and constrained by psychometric theory seem to be universally lacking. In fact, psychologists appear to have placed their faith in the psychometric integrity of their tools and pay little attention to practically applying specific information regarding reliability and validity to the decisions that have a direct impact upon the lives of their clients.

Throughout this thesis reference will be made, where applicable, to the relevant research and theoretical literature and illustrations of critical issues will be provided using simulations. Use of such simulations will facilitate both description of errors associated with current cognitive assessment and the development, demonstration and evaluation of methodologies by which these errors may be accommodated.

Chapter 2 applies reliability theory to the measurement of error at the level of the cognitive battery. Specifically this involves the discussion of reliability theory, evaluation of the impact of reliability on clinical decision making, analysis of the degree to which unreliability is evaluated at the level of the battery and recommendations regarding the use of reliability in the structure, analysis and interpretation of cognitive battery results.

Chapter 3 applies validity theory to the control of error inherent in the use of a multi-measure cognitive battery. The application of construct validation to the structuring of the cognitive battery is then outlined with recommendations for the use of empirically validated cognitive constructs. In both Chapters 2 and 3, discussion of psychometric theory will be constrained to those aspects and methodologies

specifically pertinent to the field of cognitive assessment and which are currently applied to commonly employed cognitive measures.

Chapter 4 outlines practical considerations in battery use. This involves a discussion of the degree to which factors relating to the client, the test setting and the individual tests dictate the construction and use of cognitive batteries.

Chapters 5, 6, 7 and 8 address the main aims of the thesis by proposing and evaluating a systematic method of battery structure, analysis and interpretation. Chapter 5 outlines a methodology of battery structure and analysis using composite algorithms based on classical test theorems. Chapter 6 evaluates the psychometric characteristics of the methodology by simulated investigations of the capacity of composite scores to improve reliability, formalise validated structure and answer the need for flexibility in battery construction and analysis. Chapter 7 investigates the clinical efficacy of the methodology by replication of the summary score reliability coefficients, composite intercorrelations and summary score norms for the Wide Range Achievement Test, Fourth Edition (WRAT-4), Reading Composite and the indices of the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III). Chapter 8 outlines application of the composite methodology to clinical practice based on the capacity of the proposed methodology to modify existing composite structure and facilitate the use of empirically validated alternative battery structures.

Finally, Chapter 9 involves an overview of the research, general discussion and conclusions, outlines appropriate clinical applications of the proposed method, discusses limitations of the research, and provides suggestions for future research.

CHAPTER TWO

RELIABILTY

2.1. Random Error

Error is an unavoidable aspect of measuring cognitive behaviours and one which to varying degrees confounds all statistical procedures (Chesher, 1991). Error may be systematic, due to factors which vary in close relationship to the construct of interest. Error may also be random, due to factors unrelated to the construct of interest which vary in a non-systematic fashion. Both systematic and random errors are unavoidable components of observed test scores. As discussed previously, reliability theory relates to the latter of these error sources. The term "reliability" encompasses a myriad of somewhat disparate test characteristics. However, both the mathematical bases and functional aims of reliability investigations are specifically grounded in theory (Gulliksen, 1950; Lord & Novick, 1968; Nunnally, 1978; Franzen, 2000) and are highly salient to the clinical use of both individual tests and of batteries.

Measurement or control of random error is of paramount importance to the accuracy and meaning of behavioural quantification and can be partially achieved through consideration of test reliability. In basic terms, reliability is the degree to which an observed score may be reproduced given variations in potential sources of error such as time of administration, instrument form, assessor, or variation within test items (Franzen, 2000; Gulliksen, 1950; Guilford, 1954; Lord & Novick, 1968; McDowell & Newell, 1996; Nunnally, 1978). A score is considered to be psychometrically reliable to the extent that it is robust to the impact of random error. Evaluation of the capacity of tests to measure reliably is the primary means by which random measurement error is controlled at the level of an individual test score.

Test score reliability has a fundamentally practical application as it allows individual observed scores to be interpreted within a specified degree of confidence. In other words, reliability is the psychometric characteristic of test scores which guides clinical inference to an acceptable degree of accuracy and as such is vitally important to assessors. Whenever clinical inferences are drawn from cognitive test scores, it stands to reason that the conscientious clinician would require some evaluation of the associated random measurement error. Reliability estimation ultimately holds the key to understanding the impact of such error on test usage (Franzen, 2000).

While evaluation of the reliability of individual test scores is widely accepted as the norm, the control of error at the battery level is inconsistently addressed in clinical practice. This is despite the fact that the sources of random error are likely to influence all clinical inferences, whether these are based on single test scores or on the groups of scores typical to battery administration (Chesher, 1991). Evaluation of reliability at the battery level allows for random error to be adequately understood: a fact which is acknowledged in the construction of commonly used collections of tests, such as the Wechsler scales. The reliabilities associated with the composite scores produced by such batteries are readily estimated, which is cited by fixed-battery proponents as a primary argument against flexible construction (Russell, Russell & Hill, 2005).

Given the mass and complexity of reliability literature the divide between theory and practical application provides an ongoing challenge for practicing clinicians (Charter, 2003; Sawilowsky, 2000; Thompson & Vacha-Haase, 2000). The reliability associated with any collection of test scores, however, may be determined using existing psychometric methodologies and normative data.

Application of such methods to the most prevalent form of clinical battery usage may assist clinicians to bridge the current gap between theory and practice and undoubtedly add to the psychometric strength of resulting clinical inferences. To this end the current chapter investigates reliability theory with the aim of illustrating the fundamental role of reliability in the development of accurate inferences about individuals based on batteries of cognitive tests.

2.2 Reliability as it is Commonly Applied

In classical test theory, reliability concerns the stability of the observed score as defined by the ratio of true score variance to observed score variance. As indicated earlier in formula 1.2, as the variance of the observed score increases relative to the variance of the true score, the reliability of the observed score decreases. In other words, as quantification increasingly incorporates error the resulting score is increasingly likely to vary due to this error and not to true score variation. The result of the ratio indicates the degree to which random error impacts upon the individual test score, with regards to the capacity of the score to consistently rank individuals (Thompson & Vacha-Haase, 2000). As “true score” is in effect a convenient theoretical construct (i.e., defined as the average score of an individual for infinite measurements of the current test or of infinite parallel forms; Guilford & Fruchter, 1978; Hopkins, 1998), its variance cannot be measured or calculated (Feldt & Brennan, 1989). Hence, reliability cannot be calculated but instead must be estimated by systematic variation of the possible sources of error (Franzen, 2000).

2.2.1 Classical Test Theory

Despite various insufficiencies, such as a failure to consider the potential cumulative effects of various sources of error and reliance on specific samples of test takers and test items (Thompson, 1991), classical test theory (CTT) still provides the primary theoretical basis for estimating the reliability of many commonly used cognitive tests (Franzen, 2000). As discussed in chapter one, this is due both to historical convention but also to the wide applicability of CTT due to the leniency of its assumptions. In basic terms, CTT assumes the existence of a true score, the random nature of error terms, and the linear relationship between the observed, true, and error scores, without specific assumptions regarding the functional form of true, error and observed score distributions (i.e., except for the assumption of identical distributions of observed scores from parallel tests; Feldt & Brennan, 1993; Novick, 1966). CTT is widely applied when the reliability of cognitive test scores is considered.

As some researchers have indicated, while estimates of measurement error based on CTT may produce similar conclusions as other more complex models, it bears the added advantage of simplicity (Nunnally, 1978). Certainly, estimates of error based on modern test theories (e.g., Item-Response Theory, IRT) avoid several factors which cause instability in CTT based reliability estimates. For example, independence from the sample of examinees or test items and consideration of alternative sources of error variance, constitute primary strengths of modern test theory conceptions (Embretson & Reise, 2000; Franzen, 2000). Despite this, CTT provides the theoretical basis by which the reliability of cognitive tests is most frequently considered, is perhaps better understood by clinicians, and is certainly better supported in terms of the statistical processes with which many clinicians are

familiar (Aiken, 2003; Bechger, Maris, Verstralen & Beguin, 2003; Embretson & Reise, 2000; Franzen, 2000). Charter (2003), for example, in an extensive review of reliability reporting for psychological tests found "no indication that any of the recent advances in reliability methods" were generally applied (p.298). With the aim of commenting on clinical practice as it currently occurs, CTT is the most appropriate model upon which to base the proposed methodology. While detailed discussion of modern test theory is beyond the scope of the current thesis, the degree to which the proposed methods may be modified to incorporate current advances in psychometric theory will be briefly noted where appropriate, specifically with regards to future research.

2.2.2 Estimating Reliability

Under CTT, any given observed score serves as an estimate of the true score, which would be obtained by averaging the results from administration of an infinite number of domain specific test items (Cronbach & Shavelson, 2004). Therefore, assuming the items within a test constitute only a random sample from a hypothetically infinite population of domain specific items (i.e., the Domain Sampling Model; Nunnally, 1978), the reliability coefficient may be conceptualised as the average correlation of the test, or collection of items, with all other possible tests in the domain. This may be termed r_{1j} where "1" is the test under investigation and "j" all tests in the domain. This coefficient indicates the degree to which a "common core" dictates variance and provides the best estimate of the degree to which randomly parallel forms, within the domain, would be expected to vary consistently and the degree to which they would be expected to deviate based on random error (Lord, 1955). In other words, this coefficient precisely indicates the

extent to which variance within the observed score is explained by variance within the true score.

If it could be shown that two tests were strictly parallel (i.e., having the same standard deviation, the same correlation with true scores and including no unique, systematic error) the correlation coefficient between the observed scores of these tests may be shown to directly equal the product of the ratio of true score variance to observed score variance and hence provide an expression of score reliability (i.e. see Novick, 1966, Lord & Novick, 1968, or Nunnally, 1978 for complete proof). As the correlation coefficient between parallel forms can be mathematically demonstrated to equal the reliability ratio, estimation of reliability could rely on the correlations that exist between parallel measures in the same domain (Lord & Novick, 1968; Thompson & Vacha-Haase, 2000). Random errors are not expected to correlate and such a coefficient would express the degree of true score variation present within the observed scores of each test.

Reliability coefficients are never calculated in practice, however, as strictly parallel forms are theoretical (i.e., relying on a correlation between observed and true scores) and instead are estimated. Models of reliability estimation are often based on shared variance techniques, such as correlation, and various approximations of parallel forms (Osburn, 2000). Test-retest coefficients (the correlation between scores from the same test administered on two different occasions), inter-rater coefficients (the correlation between scores from the same test determined by different assessors), and alternate form coefficients (the correlation between different but homologous test forms) all represent such an attempt (Charter, 2003). Each of these coefficients provides an expression of the random error associated with factors (i.e., time, rater, and test form, respectively) upon which approximations of parallel

measures may be based, and which may be vulnerable to differing degrees of error. Although test-retest, alternative form and inter-rater coefficients are still routinely reported, truly parallel forms are never produced by such approximations and the potential for systematic error in re-testing or with the use of alternate test forms reduce the viability of these methods as unique estimates of random error variance (Charter, 2003; Feldt & Brennan, 1993). Perhaps for this reason, coefficients of internal consistency are the most frequently reported for psychological tests and are likely to provide the best approximation of the parallel form model (Hogan, Benjamin & Brezinski, 2000; Reinhardt, 1996).

2.2.3 Internal Consistency

The psychometric literature "variously presents internal consistency as identical to reliability, as another kind of reliability. . . as a lower bound to reliability, as an approximation to reliability, or as an estimator of reliability" (Lucke, 2005, p.66). Regardless of these multiple definitions, for practical purposes, internal consistency provides the most frequently cited and perhaps the most clinically salient estimation of measurement error associated with cognitive tests. Coefficients of internal consistency indicate the degree to which random error associated with the test instrument itself is likely to impact on the observed score (Charter, 2003; Cortina, 1993; Cronbach, 1951; Cronbach & Shavelson, 2004; Schmitt, 1996; Streiner, 2003). In so doing, these coefficients reflect the random error due to the sampling of test items from the theoretical "domain" and, as such, provide coefficients of measurement error which are reliant on the parallel forms assumption. Most importantly, however, internal consistency coefficients provide the only CTT based reliability indices which do not rely upon repeated sampling thus avoiding

some aspects of systematic error variance to which other reliability coefficients are vulnerable.

Original indices of internal consistency (i.e. Spearman, 1904) relied upon the correlation coefficient between split halves of the test, frequently determined by grouping the odd and even numbered items. However, early researchers quickly identified the potential weaknesses of split-half methodology. Most relevant was that such estimates were likely to yield unstable coefficients due to the myriad of splits possible for a test of even moderate length (Brownell, 1933). Further, as reliability is fundamentally linked to the number of test items, split-half methods were likely to underestimate reliability by being based on only half the number of items in the test (Murphy & Davidshofer, 2005).

To avoid such drawbacks, several methods of internal consistency estimation were proposed based on the theoretical average of the correlation coefficients of all possible splits of test items (Reinhardt, 1996). Kuder and Richardson's (1937) famous twentieth formula (KR-20), estimated the internal consistency of a test of dichotomous items based on the number of items, proportions of positive and negative responses for each item, and the total variance of the sample.

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{N} p_i q_i}{\sigma_X^2}\right] \qquad \text{Formula 2.1}$$

The KR-20 was generalised by Cronbach (1951) to non-dichotomous test-items by the substitution of the sum of the individual item variances for the summed item difficulty statistics used in the KR-20. Conceptually, these coefficients are equivalent, however, Cronbach's formula has a much wider practical application due to its suitability for use with instruments comprised of non-dichotomous items.

Coefficient alpha is theoretically derived from the correlation coefficient between a test and a hypothetical alternate form. Calculation of the coefficient is, however, based on the average correlation between all items within a test, and in effect expresses the "average" interrelatedness of items within the scale (Schmitt, 1996). Specifically, alpha was designed to indicate the degree to which items within a measure share variance, as opposed to the variance which is specific to individual items and hence unlikely to be reflective of a common or underlying core (Cronbach, 1951). This explication readily leads to the conclusion that alpha must underestimate reliability when a test is multidimensional, which is indeed the case, as will be discussed later. Alpha provides a good estimate of internal consistency when measurements are tau-equivalent (i.e., basically the assumption of true score consistency between congeneric tests with identical units of measurement; Raykov, 1998) and provides an estimate of the lower bound of reliability for homogenous scales (Cortina, 1993; Lucke, 2005).

Computation of coefficient alpha relies on a comparison between the individual item variances and the overall variance associated with the observed score (i.e., the sum of all individual items), moderated by the total number of items within the scale, expressed in formula 2.2, as follows:

$$\alpha = \frac{K}{K-1}\left[1-\left(\frac{\sum \sigma_k^2}{\sigma_T^2}\right)\right] \qquad \text{Formula 2.2}$$

Where,"k" is the number of items on the test, $\Sigma\sigma^2_k$ is the sum all individual item variances and σ^2_T is the overall observed score variance (Cronbach, 1951).

When individual items measure the same variability between subjects (i.e., true score) then the variance associated with the overall summed observed score of the test is expected to be less than the sum of item variances. That is, all items will

incorporate unique variance which is presumed to be expressive of random error. The result of this ratio is a coefficient mathematically equal to the average coefficient obtainable for every possible split of test items (Cronbach, 1951).

Coefficient alpha is the most commonly cited coefficient of internal consistency. Cronbach and Shavelson (2004, p.392) in Cronbach's final discussion of the famous index reported "no less than 5,590" citations of the article in which the alpha formula was first presented (i.e. Cronbach, 1951). Hogan, Benjamin and Brezinski (2000) in a review of the APA-published *Directory of Unpublished Experimental Mental Measures* cite the reporting of alpha as a measure of reliability in at least two thirds of the recorded tests (p.527). Most relevant to the current study, coefficient alpha is widely denoted as the most commonly cited coefficient of reliability for cognitive tests and is used almost universally in cognitive test construction and interpretation (Brunner & Sub, 2005; Cortina, 1993; Schmitt, 1996; Streiner, 2003).

Coefficient alpha has several strengths as an estimate of internal consistency. In practical terms, it is one of the few reliability indices based on a single test administration, which is both practical and psychometrically justified. Alpha is conservative, in that it provides a lower bound of reliability under mild assumptions, and closely estimates the random error associated with the item specific variance of tests of unidimensional constructs (Cortina, 1996). Unfortunately, alpha's almost ubiquitous application may be largely attributable to the lack of subsequent coefficients. Although Cronbach intended alpha as the "first step" in a continuation of improved estimates, it seems that few viable alternatives have been proposed and coefficient alpha still constitutes the state-of-the-art in terms of CTT based coefficients of reliability (Cronbach & Shavelson, 2004).

2.3 Clinical Decision Making

Theoretical explications and practical investigations of reliability are based on the group, rather than the individuals about whom the majority of clinical decisions must be made. Reliability is the metric of random error, however, and can facilitate the use of observed test scores to evaluate cognitive behaviours of the individual with a quantifiable degree of confidence (Charter & Feldt, 2002; Feldt & Brennan, 1989; Franzen, 2000). While the quality of the method dictates the success with which random error is estimated by reliability coefficients, they should still exert a dynamic influence on clinical decision making (Franzen, 2000). The literature associated with standardised tests invariably includes some form of reliability estimate, with the presumption that clinicians are aware of the importance of reliability to the accurate and "legitimate" use of individual test scores (Charter & Feldt, 2001).

Perhaps due to the challenges of integrating psychometric theory with clinical practice (Charter & Feldt, 2002; Kline, 2000; Nunnally & Bernstein, 1994) reliability coefficients are frequently misused or not applied to clinical decisions despite their central importance and potential utility (Charter & Feldt, 2001). In practical terms, some authors believe that the application of reliability coefficients to interpretation of observed scores is hindered by the "abstract" nature of coefficients themselves (Charter & Feldt, 2002). Regardless of the reason, use of reliability coefficients in an applied setting is a difficult task which is greatly aided by the use of several indices by which error may be directly applied to understanding of the observed score. Reliability estimates, for example, provide a direct link to the true score by indicating the degree of error likely to surround the observed score. Through use of standard error indices and confidence intervals, reliability can be used to modulate the degree

of certainty with which clinicians regard test scores and subsequently exert a dynamic influence on the accurate quantification of cognitive behaviours (Franzen, 2000; Olm & Senior, 2006; Charter & Feldt, 2002).

2.3.1 Estimating the True Score

Clinicians readily acknowledge that a test score obtained by a client is a sample of their behaviour under specific test conditions (Glutting, McDermott & Stanley, 1987). Central to this understanding is the knowledge that on another occasion the client may have obtained a different score on the same test. The observed score is therefore best viewed as a sample from a distribution of possible scores which vary according to truth and error. While true score is a theoretical conception, the observed score is a sample from the distribution of possible observed scores obtained by the test taker under specific testing conditions. As such, the observed score is a tangible construct which, however, is likely to vary based on random error. However, the assumption that cognitive test scores possess exact precision is implicit when such scores are reported and interpreted by clinicians with no reference to varying degrees of score unreliability (Traub & Rowley, 1991). A perfectly reliable measure produces observed scores which are equal to their respective true score, and free from the impact of random error. When the reliability of the measure is anything less than perfect, the observed score unavoidably includes error. In this instance the reliability coefficient, provides a direct link between the observed and true scores. Reliability may be used to estimate the true score likely given the degree of error incorporated into the process of measurement. The predicted true score (PT) is derived directly from the observed score after correcting for the less than perfect reliability of the test:

$$PT = r(X - M) + M \qquad \text{Formula 2.3}$$

In this formula, "r" is the reliability coefficient, "X", the observed score and "M" the normative mean (Glutting, McDermott & Stanley, 1987). The resulting prediction of true score accounts for regression towards the mean (Stanley, 1971), therefore, deviating from the theoretical "true score". However, acknowledgement of the impact of random error on the individual observed score is achieved through use of such methodology.

The estimated true score is most commonly used to centre the confidence intervals which may be based on indices of observed score stability. It is important to note that regression towards the mean and the fundamental reliance of point estimates of the true score on the specific observed values of the observed scores upon which they are based reduces the practical benefits of using the predicted true scores as directly interpretable data points (Charter & Feldt, 2000; Feldt & Brennan, 1989). Regardless, acknowledgement of the impact of error on the observed score is conceptually appropriate and is useful in the formation of interval estimates of the true score.

2.3.2 Interval Estimates of the True Score

Reliability coefficients are most commonly employed in developing interval estimates which accommodate the likely amount of measurement error associated with observed test scores (Feldt & Brennan, 1989; Stanley, 1971). Frequently, reliability coefficients are used to compute the standard errors associated with measurement, estimation and prediction, which are then used to form confidence intervals, centred on the observed or predicted true scores. Such methodology establishes the bounds within which the true score should be expected to fall, given

the observed score and a specific degree of measurement error (Gulliksen, 1950). This methodology is not new and a certain amount of relevant debate has arisen (Dudeck, 1979; Glutting, McDermott & Stanley, 1987). Regardless, measures of average or standard error and the incorporation of such indices into confidence and tolerance intervals around test scores are cited in any text of fundamental psychometrics (e.g., Aiken, 2003; Anastasi & Urbina, 1997; Feldt & Brennan, 1989; Guilford, 1956; Murphy & Davidshofer, 2005; Nunnally & Bernstein, 1994). The indices of standard error arguably of most use in the clinical application of cognitive tests are those cited above: namely the standard error of estimate (SE_E) standard error of measurement (SE_M) and standard error of prediction (SE_P) discussed in Lord and Novick (1968).

Interpretation of individual scores, with due reference to measurement error, has traditionally occurred through the use of the SE_M (Glutting, McDermott & Stanley, 1987). Cronbach and Shavelson (2004) term the SE_M the "most important single piece of information to report regarding an instrument" (p.413), based on the capacity of this index to indicate the "uncertainty associated with each score" and the practical ease with which it may be incorporated into clinical decision making. SE_M is an expression of the error associated with estimating obtained scores from true scores (Nunnally, 1978) and can be understood as the standard deviation of scores which would be expected of a person taking a large number of randomly parallel tests (Anastasi & Urbina, 1997).

$$SE_M = \sigma\sqrt{1-r} \qquad \text{Formula 2.4}$$

SE_M is a function of the measurement error associated with the score (i.e., frequently based on internal consistency coefficients) and the normative standard deviation. Implicit in this index is the presumption that the true score (about which the index

comments) is held constant, while the given observed score is one of a number of possible observed scores (i.e., all with the same true score value) which may fall within the resulting confidence interval (Gulliksen, 1950). Instead what most test users wish to estimate is the likelihood that a given interval contains the true scores possible, given a specific value of an observed score. Constancy of the true score cannot be determined in the clinical setting, in which observed scores must suffice. Dudeck (1979) noted this misinterpretation in applied literature of the SE_M occurring, in his opinion, through a fundamental misunderstanding of standard CTT texts (i.e. Guilford, 1954; Lord & Novick, 1968).

In fact, Lord and Novick (1968) provided three formulations of standard error corresponding to three unique uses of the observed score. Given that SE_M estimates the variability of an observed score with the presumption that true score is held constant, it is correctly centred on the true score, which of course is unavailable to the clinician. Instead, where the performance of an individual is evaluated relative to a particular reference group the standard error of estimate which assumes the existence of just such a reference group is the appropriate index to use (Glutting, McDermot & Stanley, 1987). The error index suitable for establishing the bounds of true scores from any given observed score in the clinical setting, therefore, is the SE_E calculated thus:

$$SE_E = \sigma\sqrt{(r)(1-r)} \qquad \text{Formula 2.5}$$

This index is correctly centred, not on the true score, but on an estimation of the true score associated with a specific observed score value. Nunnally (1978) and Dudeck (1979) both stressed the asymmetrical placing of confidence intervals based on SE_M around the observed score, which instead must be symmetrically placed around the estimated value of the true score computed using equation 2.3. A

confidence interval surrounding the predicted true score is calculated using formula 2.6, as follows:

$$PT \pm (z) * \left(\sigma\sqrt{(r)(1-r)}\right) \qquad \text{Formula 2.6}$$

"PT" is the predicted true score associated with a specific observed score and "z" is the appropriate value of the unit normal deviate corresponding to the desired level of confidence.

Finally, when the prediction implicit includes the determination of a re-test score, the appropriate coefficient is the SE_P calculated thus:

$$SE_P = \sigma\sqrt{1-r^2} \qquad \text{Formula 2.7}$$

The SE_P, centres on a predicted observed score, and indicates the interval within which a subsequent observed score on the same test is likely to fall, given measurement error and regression to the mean. Re-test scores on the same measure falling within the resulting confidence interval are not considered indicative of systematic changes in true score performance. Rather such an observed re-test score is said to deviate based simply on the instability of the test and not the influence of a systematic source of variation, such as actual change in true score.

These indices of error provide a means of directly applying reliability to clinical decision making and as such seem essential tools for assessment clinicians. As discussed in the current section, the methodological tools necessary to actively use reliability when drawing clinical inferences are both readily available in the psychometric literature and simply applied.

2.3.3 The Impact of Reliability on Clinical Decision-Making

Reliability based confidence intervals provide a valuable means of avoiding errors of overconfidence likely when test scores incorporate error. The impact of

unreliability on clinical decision making can be emphatically demonstrated in terms of the quantifiable error associated with clinical decisions made using cut-off scores.

Charter and Feldt (2001), in a simulated example, demonstrated that due to the impact on standard deviation, reliability moderated the cut-off score required for accurate clinical decision making. In this study, the rate with which clinicians would make both false negative (FN: misclassification of an impaired examinee as not impaired) and false positive (FP: misclassification of a "normal" examinee as impaired) errors was directly influenced by the extent to which observed scores incorporated measurement error. Specifically, when a cut-off score for impairment was specified the relationship of the individual's true score to this arbitrary point was fundamentally confounded by the incorporation of random error. The probability associated with incorrect classification of individuals based on the relationship of observed test scores to a given cut-off (i.e., in terms of both FN and FP conclusions) was directly modified by the reliability of the observed scores. In this study, the authors demonstrated the use of reliability coefficients to moderate cut-off scores based on the likelihood that the observed scores to which they will be compared are likely to be unreliable.

While not every clinician would be willing to employ the methodology indicated in the above study, it does provide emphatic evidence for the fundamental relevance of score reliability to accurate and error free clinical decision making. Reliability directly influences the probability of successful decision making using observed test scores and without consideration of score reliability the rates of decision making errors are altered. This highlights the necessity of considering reliability at the level of the individual and test score, rather than at the level of the test.

2.4 Cautions for Interpretation of Reliability Coefficients

Use of reliability and internal consistency coefficients in clinical decision making is, however, vulnerable to several errors. In fact, several authors have cautioned about practical misuse of reliability arising from a misunderstanding of the strengths, weaknesses and specific characteristics of coefficients. Alpha, for example, is frequently misapplied leading some authors to caution against its almost ubiquitous application (Cronbach & Shavelson, 2004; Cortina, 1993; Streiner, 2003; Reinhardt, 1996). Several cautions are universally applicable to the clinical interpretation of CTT based reliability coefficients and should be considered with the aim of better utilising reliability.

2.4.1 Recommended Reliability Levels

The literature is replete with recommendations of the minimum required internal consistency coefficient. A well accepted convention within the literature is that reliability coefficients must be high and rarely less than r = .85. Table 2.1 below summarises recommendations in the psychometric literature regarding the strength of reliability necessary for stable and accurate test based inference. As indicated in the table, authorities vary in their recommendations, while concurring that coefficients must be at least r =.7 or over and optimally nearer to r = .8 or r = .9.

Table 2.1

Summary of Recommended Reliability Levels in Psychometric Literature

Source	Standard of Reliability
Aiken (1991)	r = .85 or higher when scores are to be used to make clinical decisions
Cicchetti (1994)	r <.7 is unacceptable, and that test developers should aim for .7 - .8, and optimally >.9
Guilford, (1950)	r = .9 is necessary to measure accurately
Gregory (1999)	r = .9 when a test is used clinically
Guilford & Fruchter (1978)	r = .9 for clinical use
Hopkins, Stanley & Hopkins (1990)	r = .9
Kelly (1927)	r = .94 is necessary to evaluate individual performances
Kline (2000)	"reliabilities should ideally be high around .9, especially for ability tests. " (p.13)
Nunnally (1978)	r = 1.00 when test are clinically applied "it is frightening to think that any measurement error is permitted" (p.246)

Table 2.1 (continued)

Source	Standard of Reliability
Nunnally & Bernstein (1994)	r = .8 is appropriate for research purposes "If important decisions are made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard" (p.265)
Rosenthal & Rosnow (1991)	r = .85 for clinical decisions
Salvia & Ysseldyke (1988)	r = .9
Sternberg (1994)	"For diagnostic and screening tests one will be very uncomfortable with reliability estimates below .8 and prefer the reliability to be .9 or higher" (p.954)
Weiner & Stewart (1984)	r = .85 for clinical decision making

Some acceptance of imperfect reliability seems unavoidable in test usage. It is not, however, unwarranted to reiterate that given the likely impact of cognitive testing results "it is frightening to think that any measurement error is permitted" (Nunnally, 1978, p.246).

The reliability coefficients for individual tests presented in the literature are on average lower than the recommendations outlined in Table 2.1. For example, in a review of over 937 published reliabilities for clinical, personality, vocational, neuropsychological, intelligence and educational tests undertaken by Charter (2003)

cited reliabilities were often somewhat lower than those recommended in the psychometric literature. In this study, the average internal consistency coefficient for the thirty-two neuropsychological tests surveyed was r = .82 which varied by a standard deviation of .16 (see Table 2.2, derived from Charter, 2003, for average coefficients for all categories of tests surveyed in this study).

Table 2 2

Mean Reliability Coefficients for Psychological Tests

Coefficient	Mean Reliability	SD	Min/max	N
KR-20	.83	.10	.61-.98	62
Alpha	.81	.16	.23-.98	97
Alternate Form	.84	.10	.40-.96	40
Split-half	.83	.12	.35-.98	126
Retest	.79	.13	.17-.99	439
Interjudge	.86	.14	.36-.99	84

Even when reliability levels are not optimal the clinician would do well to be aware of the extent to which cognitive tests demonstrate suitable levels of reliability prior to incorporation into a testing battery and, most importantly, when the scores from tests are interpreted.

2.4.2 Theory and Cause of Measurement Error

No single reliability coefficient can suffice to describe the impact of "error in general" as reliability coefficients are used to determine the impact of various sources of error. Instead multiple classes of reliability estimates are needed to comprehensively conceptualise the impact of random error on an observed score. In terms of an actual cognitive test "the reliability" may refer to the likely vulnerability of scores on the test to errors attributable to the test items, test form, test raters or administration times. All models of reliability do not consider the same sources of random error and thus cannot be presumed to yield estimates which are equally useful to any given application of the test score. In fact, test score reliability is likely to be fundamentally affected by the measurement theory by which the estimate is produced. The theory of reliability relating to a particular coefficient is highly relevant and must be taken into consideration when reliability is incorporated into practical clinical decision making (Sawilowsky, 2000; Thompson & Vacha-Haase, 2000).

This distinction becomes most obvious when considering the specific vulnerabilities of the particular reliability coefficient employed. Coefficient alpha, for example, provides an acceptable estimate of internal consistency when scale indicators are tau-equivalent (or essentially so) and when error terms are uncorrelated (Brunner & Sub, 2005). Deviation from these conditions weakens alpha as an estimate of internal consistency. Alpha does not ensure stability of test scores over time, does not indicate the equivalence of scores across alternative forms and, on a distinct but related note, is often incorrectly interpreted as a measure of unidimensionality (Crocker & Algina, 1986; Schmitt, 1996). The universal application of alpha in the testing literature and its ubiquitous appearance in test

manuals does not ensure its' universal applicability to clinical decision making. Instead, several authors have cautioned that interpretation of coefficient alpha as a measure of reliability must be undertaken with reference to the specific vulnerabilities of the coefficient (Cronbach & Shavelson, 2004; Streiner, 2003).

2.4.2 Specificity to Sample

Reliability, grounded on CTT tenets, is not a property of tests, but a property of test scores and as such is fundamentally sample specific (Guilford, 1956; Novick, 1966; Vacha-Haase, Kogan & Thompson, 2000). In other words, any given reliability coefficient is a characteristic of scores on a test derived from a particular sample, and not a characteristic of the actual instrument itself (Crocker & Algina, 1986; Fan & Yin, 2003; Thompson & Vacha-Haase, 2000; Yin & Fan, 2000). Factors intrinsic to the individual sample participants, such as heterogeneity regarding the trait being measured (i.e., group variance) and overall level of the trait (i.e., group average), impact upon the rank ordering within the individual sample and hence affect reliability estimates (Fan & Yin, 2003; Gulliksen, 1987; Linn & Gronlund, 1995; Thompson & Vacha-Haase, 2000).

For this reason, psychometricians argue the need for clearer descriptions of samples within the literature (Vacha-Haase, Kogan & Thompson, 2000; Vacha-Haase, Ness, Nilsson & Reetz, 1999). While arguments are often made in terms of the failure of researchers to adequately report sample specific reliability coefficients, the cautions are equally applicable to the use of reliability estimates in the field of clinical testing. Regardless of application, the onus is clearly on test users to determine the suitability of a particular reliability estimate for the individual situation to which it will be applied (Crocker & Algina, 1986).

Additionally, estimates of reliability are themselves vulnerable to measurement error (Charter, 2003; Vacha-Haase, Kogan & Thompson, 2000) and several authors have specified the sample size necessary to achieve a sufficiently stable estimate of reliability. It is somewhat ironic that at the same time that learned practitioners can emphasise and recognise the importance of reliability in test selection and test score interpretation, they tend to treat reliability coefficients as test-specific constants. In reality, reliability coefficients are estimates with associated confidence intervals. Contemplation of the computation of confidence intervals for reliability estimates which are then utilised in determining confidence intervals to be placed around observed scores goes a long way to suggesting the barriers to incorporating psychometric considerations into clinical decision-making.

Nunnally (1978) stated that a representative sample of at least 300 participants is required to ensure sampling errors will present only "minor consideration" to a coefficient. More recently, Charter (1999) indicated a sample size of 400 participants to achieve a stable and precise estimate of test reliability. A review, conducted by this author of over 6,322 reliability coefficients provided in research and testing literature indicated the use of substantially smaller samples. For example, in examining 1,708 neuropsychological studies, the mean reported sample size was just over 92, substantially less than the recommended 400 (see Table 2.3, derived from Charter, 1999). If reliability estimates are derived from a sufficiently large sample they are likely to be less vulnerable to measurement errors. However, this does little to alleviate the impact of sample variance or overall trait levels on coefficients.

Table 2.3

Samples Sizes used to Produce Reliability Estimates

	Number of Studies	Mean Sample Size	SD Sample Size
Type of Test			
Neuropsychological	1,708	92.24	86.90
Intelligence	794	192.71	159.52
Personality	1,320	234.57	388.69
Type of Coefficient			
Internal Consistency	2,673	344.85	635.56
Generalizability	258	81.84	33.1
Interrater	217	37.00	40.22
Retest	3,174	81.70	67.70

As with other coefficients of test score reliability, alpha is not a fixed property of the scale, instead it is a property of the test scores for a specific sample of test takers. An estimate of coefficient alpha for a specific test may therefore be expected to differ based on attributes of the sample from which it is derived. Heterogeneity or homogeneity in the construct of interest and the resulting total and true score variances are particularly likely to impact on total score variance and hence on estimates of coefficient alpha (Fan & Yin, 2003). Sample heterogeneity, for example, results in larger total score variance and subsequently higher reliability

estimates. Conversely, homogeneity decreases variance within the sample and thus may lower estimates of reliability. A scale with a large coefficient alpha in a group which differs widely in terms of observed scores may have substantially lowered coefficient alpha in a group which is more similar in terms of the construct of interest.

This becomes perhaps most salient in the comparison between alpha coefficients derived from clinical and normative groups. In fact, misunderstanding of this fundamental characteristic of reliability estimates may lead to clinical misinterpretation (Streiner, 2003). For example, the reliability of a test score in a particular setting may be substantially higher than in the study from which the reliability coefficient is being taken. In such circumstances the clinician may be utilising a much greater margin of error than is necessary along with the consequent toll this takes upon accurate decision-making. Of course the advantages attendant on use of a reliability coefficient based on sample which is highly representative of any given client are negated by factors such as sample size. The development of coefficients specific to a particular clinical practice presents a potential solution, though one which is admittedly logistically difficult.

2.4.3 Reliance on Test

"The test score has practical significance insofar as it is representative of the individual's ability to respond to all of the tasks in the universe which it undertakes to sample" (Stanley, 1971, p.407). Any given test is simply a sample of such tasks and as such does not constitute the "universe" of possible tasks. Thus, while the reliability coefficient is not specific to the instrument, it is fundamentally test bound and specific characteristics of the test may impact substantially upon reliability

estimates. Most fundamentally, reliability coefficients vary based on the number of individual test items (Lord & Novick, 1968). On a related note, while the mathematical properties of reliability models remain constant regardless of the factorial composition of the domains tapped by test items (Nunnally, 1978), reliability estimates are fundamentally vulnerable to the construct dimensionality of test items. This is particularly true of coefficient alpha.

Items from a factorially diverse instrument may still share sufficiently high intercorrelations to ensure a substantial estimate of internal consistency. This is particularly likely if the test is long and coefficient alpha is used (Nunnally, 1978). While it may seem reasonable to assume that randomly sampled items from the same domain are highly (and consistently) correlated, this is not always the case and internal consistency amongst test items as a whole is not sufficient to indicate that such items share a common factor: in fact several factors may be present within test items. In other words, reliability is a necessary but not sufficient requirement for construct validity (Streiner, 2003).

Coefficient alpha is incorrectly assumed to indicate the degree of first factor saturation or homogeneity present amongst test items and some authors have highlighted the misapplication of alpha in this respect (Cortina, 1993; Cronbach & Shavelson, 2004). While a substantial coefficient alpha indicates that test items are interrelated (i.e., that is they contain very little unique variance), and while alpha will increase with item intercorrelation and decrease with multidimensionality, it does not guarantee the homogeneity of the inter-item correlations or their unidimensionality (Cortina, 1993; Schmitt, 1996; Streiner, 2003). Cortina (1993) for example, demonstrated that a scale consisting of up to three orthogonal subscales could still produce test scores with alpha estimates greater than .60, when the test included

sufficient items. In other words, alpha may be high even when a test is strongly multidimensional and has some very low item intercorrelations. Alpha does not guarantee item homogeneity, though the two concepts are complementary, and instead the dimensionality of the scale must be established through other means, for example factor analysis (Cortina, 1993).

As an aside, this misunderstanding seems related to Cronbach and Meehl's (1955) original discussion of construct validity, in which the authors outlined the complementary role of coefficients of internal consistency in establishing construct validity. These authors indicated that construct validity may be evaluated in terms of the internal structure of test items. While coefficient alpha does indicate internal structure, it does not rely on consistent correlations between test items and, as discussed, can be elevated even when some item intercorrelations are extremely low, such as those necessary for the existence of orthogonal factors.

On a related and perhaps more relevant note, given its general sensitivity to measure dimensionality, several authors stress the unsuitability of coefficient alpha as a reliability estimate for multi-dimensional measures (Cortina, 1993; DeVellis, 2003; Streiner, 2003). Streiner (2003) for example, argues that estimation of a high (i.e., >.9) coefficient alpha in a multi-dimensional test may simply indicate redundancy between test items which are related. Brunner and Sub (2005) indicate that the violations of tau-equivalence and correlation of error terms likely in multidimensional measures threaten the appropriateness of coefficient alpha as a means of understanding reliability for such scales. While Cronbach (1951) originally argued the applicability of coefficient alpha to multi-dimensional instruments (along with specifically rebutting the previously held assumption that "items on a test were unidimensional") he questioned the clinical relevance of such application given the

complexity of the resulting interpretability. Various other authors have argued the necessity of test unidimensionality, on the grounds of “psychological sense” (e.g., Hattie, 1985).

This discussion simply scratches the surface of a fundamentally salient issue: that is, the psychometric interpretability of multidimensional measures. A good many highly useful and clinically meaningful scales are designed to tap constructs with a certain degree of heterogeneity, a common example being speeded tests which reflect speed in addition to the specific cognitive ability of interest (Brunner & Sub, 2005; Cronbach & Shavelson, 2004; Lucke, 2005). In fact, tasks of increasing cognitive complexity are increasingly unlikely to be homogenous in nature (Nichols & Kuehl, 1999) and several authors convincingly argue the need for psychometric measures to be heterogeneous in order to maintain essential properties of validity (Lucke, 2005; Raykob & Shrout, 2002).

Given the heterogeneous nature of cognitive constructs, valid cognitive tests are likely to be heterogeneous. As Brunner and Sub (2005) argue, very few reliability analyses "take the multidimensional structure of empirical data into account". Thus, the weakness of coefficient alpha to indicate internal consistency in a lengthy, multidimensional test is an attribute of many alternative, and perhaps less useful, reliability coefficients. Arguably, finding a substantial coefficient alpha in the presence of significant heterogeneity amongst test items provides little evidence that the coefficient is unusable as an estimate of reliability. It does, however, add to the complexity and ambiguity of the resulting test score in terms of error (Cortina, 1996). For example, it is important to note that when multidimensional tests are evaluated, coefficient alpha is likely to produce greater standard errors and underestimate reliability (Cronbach, 1951; Streiner, 2003).

The reliability coefficients which accompany many cognitive tests are likely to be considered as inflexible constants and to be used with little thought as to their accuracy, the method with which they were derived, or the appropriateness of the sample with regard to their particular case. This highlights a fundamental problem with the common clinical application of reliability coefficients. Coefficients must be used with some reference to relevant characteristics of the sample from which they were derived, the likely stability of the estimate, the model upon which estimation was based, the error source being evaluated and the construct validity, or at least dimensionality, of the test instrument itself. These concerns, however, in no way alleviate the need to consider reliability coefficients as a necessary aspect of test-score based inferences (Franzen, 2000). While coefficient alpha will imperfectly estimate internal consistency in many instances, for example, it is still the most widely available coefficient of reliability for cognitive tests and provides the clinician with more information about random measurement error than an observed score considered in isolation from reliability estimates. Better than ignoring the presence of random error altogether, is the use of somewhat flawed coefficients with due consideration of their vulnerability. In fact, the most salient fault in clinical use of reliability coefficients is the failure to apply reliability to the drawing of clinical inferences.

2.5 Reliability at the Battery Level

In assessing the cognition of adults, clinicians are highly likely to include more than one test of any given construct or ability. Table 2.4 (derived from Olm & Senior, 2006) for example, lists several of the cognitive tests which clinicians may use in the measurement of cognitive functioning, organised by the cognitive abilities with which they are routinely associated. As indicated, several tests of each "domain" are available and clinicians are likely to draw conclusions about domain-based functioning using several domain-related measures (Vanderploeg, 2000).

Table 2. 4

Commonly Used Tests

Construct	Tests
Verbal Comprehension/ Word Knowledge	Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) Verbal Comprehension Index, Shipley Institute of Living Scale (SILS)-Vocabulary, Speed and Capacity of Language Processing test (SCOLP)-Spot-the-word (STW), Wide Range Achievement Test, Third and Fourth Edition (WRAT-3; WRAT-4) – Reading
Verbal Fluency	Controlled Oral Word Association Test (COWAT), Animal Naming (ANIMALS), Stroop Neuropsychological Screening Test (STROOP)
Visual Processing Abilities	WAIS-III Perceptual Orientation Index (POI), Visual Forms Discrimination (VFD), Judgement of Line Orientation (JLO), Hooper Visual Orientation Test (HVOT), Rey Complex Figure Test (RCF)– Copy
Attention/Concentration	WAIS-III Working Memory Index (WMI), Wechsler Memory Scale, Third Edition (WMS-III) WMI, STROOP
Speed of Processing	WAIS-III Processing Speed Index (PSI), STROOP, COWAT, SCOLP-Speed of Comprehension (SOC), ANIMALS, Symbol Digit Modalities Test (SDMT), Trail Making Test (TMT) - Trial A

Table 2.4 (continued)

Construct	Tests
Naming	Boston Naming Test (BNT)
Memory	WMS-III, Rey Auditory and Verbal Learning Test (RAVLT), RCF
Visual Scanning	SDMT, TMT – Trial A and Trial B
Visual Tracking	Rey Tangled Lines Test (RTLT)
Problem Solving/Mental Flexibility	Wisconsin Card Sort Test (WCST), SILS – Abstraction, TMT – Trial B, Booklet Category Test (BCT)
Psychosocial Functioning	Minnesota Multiphasic Personality Inventory, Second Edition (MMPI-2), PAI, PCL, CRS
Motivation, Poor Effort, Malingering	Rey 15-Item Test, WAIS-III, WMS-III, RAVLT, RCF, WCST, Test of Memory Malingering (TOMM), Symptoms Validity Test (SVT), MMPI-2

Scores, from different tests, which relate to the same domain, are likely to be interpreted not only in terms of their individual qualities, but in terms of the qualities they share. Clinicians may be interested in the magnitude of similarities or discrepancies between such scores or in their shared relationship to test scores of another construct. Such implicit score combination is simply a more complex form of the behavioural quantification that occurs when using individual tests. Inferences about the cognitive functioning of an individual drawn either from single test scores or informally combined "composites" of tests are fundamentally equivalent in terms of their underlying inference and, most importantly, in terms of their vulnerability to the impact of random error.

In fact, when clinicians cluster tests in their heads into domain related groupings, they assume more than just the assertion that the tests measure the same underlying construct. Instead, such score combination presupposes that measures have perfect reliability and high intercorrelations. In such a situation measures are given equal weighting in terms of their capacity to measure construct-relevant behaviours or this capacity may be attributed based on arbitrary test characteristics rather than empirically evaluated qualities of the measures. When tests are anything short of completely reliable, and instead vary in terms of both reliability and inter-relation, clinicians must recognise these differing capacities when using the grouped measures to draw inference. The unique reliabilities of domain-specific measures and how they interrelate will vary from test combination to test combination. To assume that these issues do not matter reflects 'worst practice' and is unfortunately all too frequently observed both implicitly (through assumptions) and explicitly (e.g., pro-rating and test substitution) in current assessment practices. In fact, the process of drawing inferences "formally" from individual test scores and "informally" from

unquantified comparisons between test scores arguably differs only in terms of the degree to which psychometric theory is applied and fundamental differences in reliability and intercorrelation are acknowledged.

This issue may be considered in light of a domain-sampling model (Lord & Novick, 1968). Items which are intended to tap a particular construct are arguably drawn from the same domain, regardless of whether they are collected into a single test or encompass several tests. In other words, items which are intended to tap the same domain may be considered as items selected from the same hypothetical "universe" of possible construct-related measures (Nunnally, 1978). It follows that such items may be analysed and interpreted in relation to each other and that, presuming items are valid measures of the intended construct, changes in the underlying domain are likely to affect these items in essentially similar ways.

Measurement in this case is conducted with the aim of estimating the individual's "true" score on the domain. Such measurement raises questions of reliability, as surely as does the use of test items which are formally combined into a single measure. The sources of random error remain, and in fact may increase, when individual test scores are combined for clinical interpretation, albeit informally. The need to understand error at the battery level is as important as the well recognized need to evaluate the reliability of individual tests and various authors have commented at length on the drawback of failing to evaluate reliability at the battery level (Russell, Russell & Hill, 2005). As Murphy and Davidshofer (2005) state domain measurement using a single test or a combination of tests is conceptually equivalent and "adding together scores on several highly correlated tests achieves exactly the same objective as adding together scores on several positively correlated test items to form a single test score" (p.147). Evaluation of the reliability of the

resulting item combination is patently necessary in both instances, a fact which is increasingly recognised by assessment specialists (Miller & Rohling, 2001).

When cognitive measures are developed the reliability of the entire measure must and is likely to be evaluated: this is not only common practice but a requirement of professional codes of conduct in psychology (Australian Psychological Society, 2007). Informally combining scores from tests of the same construct for the purposes of drawing clinical conclusions about the domain-based functioning of an individual is theoretically akin to developing just such a cognitive instrument. Any valid score combinations could profitably be subject to the same scrutiny, in terms of capacity to rank individuals consistently, as individual observed scores. The stability of domain-based inferences is rarely considered, however, when semi-flexible cognitive batteries are used clinically. This does not occur, however, through lack of an appropriate psychometric methodology. In fact, the overall reliability of groups of test scores of the same domain can be readily evaluated using composite methodology available in basic psychometric texts (i.e. Murphy & Davidshofer, 2005) and outlined comprehensively in the literature (Atkinson, 1991; Ghiselli, 1964; Guilford, 1954; Lord & Novick, 1968; Mosier, 1943; Nunnally, 1978; Schretlen, Benedict & Bobholz, 1994; Tellegen & Briggs, 1967; Wechsler, 1981; Wechsler, 1997a). Such methodology is comprehensively outlined in chapter five of this thesis.

The failure to consider reliability at the level of the semi-flexible battery is likely to impact considerably on clinical decisions. On the other hand, evaluation of the reliability associated with the combined analysis of several domain-specific tests provides several advantages to clinicians. Specifically such evaluation aids in test selection (i.e., the number of tests required to reliably evaluate a construct), test

combination and analysis (i.e., increasing reliability of the overall composite group provides a rationale for grouping tests), and test interpretation (i.e., the reliability of test combinations indicates the degree of confidence that is warranted in the measurement) and provides clinicians with necessary information regarding the impact of random measurement error on battery scores.

2.6 Conclusions

Reliability relates to what is the first fundamental consideration in cognitive battery usage, that of score stability and the influence of random measurement error. Without direct application of reliability theory to clinical inference, the measurement of real-world cognitive behaviours with observed test scores is fundamentally flawed.

The impact of random error is to occlude clear measurement. A given test score is rarely exclusively comprised of true score. Instead random error prevents the precision of the measurement and estimation of this error is necessary to dictate the extent to which the observed score would be replicable and the degree of confidence placed in the measurement. A stable reliability coefficient, derived to indicate the impact of a relevant source of error, should play an integral part in clinical inference. If this does not occur, the clinician can never state with certainty the degree to which random error is likely to impact upon clinician conclusions. While this argument is delineated clearly in the literature in terms of individual scores, few attempts are made to capitalise on the information provided by reliability estimates when semi-flexible batteries are used. This is despite that fact that such informal groupings of scores may be usefully evaluated in terms of their shared reliability using readily available methods. As Feldt and Brennan (1993) sagely

observed “lack of knowledge about measurement error does not remove it from any set of scores” (p.105) and quantification is, after all, only useful to the extent that is does not occlude the truth.

All reliability estimates are not equally reliable and instead may be evaluated in terms of their stability, the source of the error variance upon which they comment, and the match between the sample from which they were derived and the individual test taker. In terms of research studies, psychometricians caution against inappropriate generalisation of reliability coefficients and argue that the appropriate estimate of reliability for a particular test usage is the coefficient associated with that specific use (Vacha-Haase, Henson & Caruso, 2002). Reliability coefficients derived in differing clinical settings, or found within test manuals and in previous studies are not universally applicable, but highly dependant upon the sample from which they were derived (Crocker & Algina, 1986; Vacha-Haase, Kogan & Thompson, 2000).

The lack of appropriate reliability coefficients may prompt the clinician to derive estimates from their own clinical sample using readily available procedures and simple spreadsheet commands (i.e. see Vacha-Haase, Henson & Caruso, 2002). The most likely solution, however, is that clinicians seek reliability estimates in the literature derived from samples similar to their own patient populations. Specifically, internal consistency coefficients obtained from a large scale, clinical sample (i.e., with greater than 400 participants; Charter, 1999) from a clinical setting similar to that in which the test is used are likely to provide accurate and stable estimates of the reliability of individual test scores in measuring individual cognitive behaviours.

Most importantly, however, reliability could usefully be considered in any instance in which score-based inferences occur. This is regardless of the formality of

such inferences or the number of individual observed scores upon which they are based. Use of existing methodology to determine the reliability associated with a group of tests of a given construct could substantially improve understanding of random error at the semi-flexible battery level.

Evaluation of random measurement error, however essential, is fundamentally restricted to a specific and narrow role in psychometric theory (Nunnally, 1978). As Wilson (1998) somewhat whimsically comments reliability simply expresses a test's capacity to "reproduce itself" regardless of any desire on part of clinicians and researchers for reliability to offer insight into meaning. In fact, the question of "what tests actually measure" is perhaps the most pertinent one and application of reliability at the battery level relies fundamentally on the existence of a stable battery structure, which may only be achieved through the empirical determination of clear and distinct cognitive constructs. This issue is discussed in the next chapter.

CHAPTER THREE

STRUCTURING THE BATTERY THROUGH VALIDITY

3.1. Systematic Error

Measurement of cognition using tests should be conducted with the accuracy necessary to draw correct inferences about real-world behaviours and abilities. The specific characteristics of a cognitive ability are unlikely to be directly observed, and instead are inferred from their manifestations on construct-relevant test items. As discussed in chapter one, this process is susceptible to both random and systematic errors. While random error may be more or less directly understood through the process of reliability, systematic error is difficult to dissociate from the true score with which it consistently varies. In fact, under CTT, systematic errors may be understood only indirectly through evaluation of the nature of true score. Despite the difficulty in accomplishing this, investigation of the meaning, or validity, associated with the systematic component of observed scores is a more fundamental task than control or measurement of random error (Anastasi & Urbina, 1997; Hogan, 2003). Several authors argue that in the absence of established validity other psychometric indicators become "relatively inconsequential" (Feldt & Brennan, 2004; Hogan & Agnello, 2004; Nunnally, 1978). Validation, in fact, pertains to the very essence of assessment, which is building an accurate and representative model of the real world.

The mathematical estimation of reliability provides little or no insight into the nature of true scores. Specifically, reliability does not add information about the nature of the traits or abilities measured by test items, the degree to which test items comprehensively sample relevant facets of those traits, or the degree to which scores from the test facilitate correct inferences about the test taker's "true" abilities. Even the most reliable score may measure an entirely meaningless construct (Feldt &

Brennan, 1989). While reliability facilitates stable observation, it does not ensure correct clinical inference.

In CTT terms, the true score is the component of the observed score which is consistent at a given level of the underlying trait. This score, by definition, is free from random error, but is unlikely to entirely reflect the construct it purports to measure (Streiner, 2003). As discussed earlier, the inexact nature of psychological constructs much reduces the likelihood that the "true" score component of a given observed score is an error free measure of a unidimensional construct. Instead, the observed score may be expected to vary partially due to other factors which share similar patterns of variance. "True score" incorporates a covert degree of error caused by factors such as additional cognitive abilities which may be required to complete a test, or systematic aspects of the test setting, assessor or individual (Cimino, 2000; Franzen, 2000). Any sources of error which produce systematic variance are indistinguishable from the true score using CTT based reliability estimates: a fact which essentially impacts on the veracity of clinical inferences.

Invalidity reduces the meaningfulness of behavioural quantification as surely as does random error variance and ultimately results in erroneous clinical inferences. Psychometric validity provides the model by which test score meaning may be understood and controlled. What is meant by "validity", however, is far from clear even in the psychometric literature and has historically been the topic of some debate (Cattell, 1964; Cronbach & Meehl, 1955; Kane, 2001).

CTT conceptions of reliability are readily expressed in mathematical terminology perhaps facilitating the wealth of research literature devoted to these topics. Despite this, as discussed in chapter two, reliability theory is inadequately applied in clinical practice, a tendency which is perhaps even more pronounced in

terms of validity. Unlike psychometric reliability, the validity of test scores is less specifically definable and encompasses a broad field of often disparate investigations. As Kane (2001) eloquently comments “it is notoriously difficult to pin down the interpretation. . .of an observation” (p.339); which is, in effect, the difficult task undertaken in the process of psychometric validation. While some understanding of the extent to which scores are meaningful or “valid" is clearly needed when clinical inferences are drawn from individual or grouped test scores, application of the theory to clinical practice again presents significant challenges to the real-world use of cognitive tests (Bigler, 2007; Russell, 2007).

Nonetheless, some estimation of psychometric validity is a well accepted requirement for the production of clinically useful results from individual tests (Anastasi & Urbina, 1997; Lezak, 1995; Lezak, Howieson & Loring, 2004; Mitrushina, Boone & D’Elia, 2005; Murphy & Davidshofer, 2005; Nunnally, 1978; Spreen & Strauss, 1998) and validity is most essential when batteries of test scores must be interpreted according to meaningful cognitive domains. Bridging the gap between theory and practice, however, is still incomplete when semi-flexible cognitive batteries are used clinically. This warrants discussion, as validation pertains to the second fundamental consideration of battery construction, that of structure.

3.2 Validity Theory

In basic terms, “validity” refers to the meaningfulness of scores obtained via an instrument or test (Cronbach & Meehl, 1955; Foster & Cone, 1995). A valid test yields scores which may be meaningfully interpreted for the required purpose. In terms of cognitive tests, validity often pertains to the score’s usefulness in

generalising to actual cognitive abilities. In broad terms, the validation process occurs to determine the behavioural inferences which may be appropriately made from test scores, and must vary in focus as different behavioural inferences are required of the instrument (Kane, 2001). As such, validation is a continual process which provides the test user with empirical evidence of the acceptability of inferences (Cronbach & Meehl, 1955; Dunnette & Borman, 1979; Franzen, 2000).

The validity associated with the use of a test is not a unitary concept and instead may require the empirical evaluation of the usefulness of the test for a myriad of possible purposes (Dunnette & Borman, 1979; Foster & Cone, 1995; Franzen, 2000; Murphy & Davidshofer, 2005). Strictly speaking, however, all validation pertains to the inferences drawn from the existing or eventual observed test scores and not to the actual instrument (Franzen, 2000; Hogan, 2003). Validity fundamentally relates to the difficult task of generalising past the testing situation rather than ensuring the instruments used are foolproof. For this reason, the choice of the most correct validation procedure directly depends on the subsequent inferences for which test data are required (Foster & Cone, 1995).

3.2.1 Historical Perspective

Historically, individual tests have been evaluated in terms of their ability to adequately represent a specific population of content (content validity), the degree to which they are consistent with hypothetical constructs (construct validity), their relationship to other measures of the same construct (criterion validity), their facilitation of tasks such as the prediction of future functioning (predictive validity), their contribution to categorisation of individuals (diagnostic validity), their clinical meaningfulness (ecological or treatment validity), or whether they appear to measure

what they purport to (face validity), among other tasks (Anastasi & Urbina, 1997; Foster & Cone, 1995; Kane, 2001; Kane, 1992; Murphy & Davidshofer, 2005; Nunnally, 1978).

Early validation was largely criterion-oriented (Cronbach & Meehl, 1955). Observed scores from the test were evaluated in terms of their ability to predict or approximate a criterion measure, frequently by use of correlation and related shared variance techniques (Kane, 2001). Implicit in criterion-oriented validity, however, is the assumption that the criterion provides the operational definition of the construct of interest. In many instances the lack of a valid or relevant criterion forces the use of indirect measures and in the absence of a suitable criterion alternative validation techniques were necessary. For example, the model of construct validity was proposed for validation of "attributes for which there is no adequate criterion" with authorities such as Cronbach and Meehl (1955) declaring criterion based validation to be "obsolete".

3.2.2 Current Validation

Current test validation techniques generally require the identification of the intended future uses of test scores, an explicit statement of the assumptions and inferences inherent to the use of such scores, and empirical and logical evaluation of the reasonableness of the intended inferences, without clear adherence to a specific "kind" of validity (Kane, 2001). The complexity of the validation procedure increases proportionally with the abstractness of the relevant construct (Nunnally, 1978). As Kane (2001), states "the validation of a spelling test as a measure of skill in spelling words in some domain of words need not involve the same level of effort, or the same kinds of evidence, as the validation of a theoretical construct embedded

deep in a complex theory." (p.332). Thus, while highly concrete variables require straight forward validation and are frequently definable in terms of test content or their capacity to predict a relevant criterion, increasingly abstract variables require increasingly complex validation procedures to define the theoretical "constructs" which underlie test items. At the level of the individual cognitive test, validation procedures may pertain to relatively straight-forward inferences, such as the degree to which test items are representative of a particular domain or the predictive abilities of the test. Most commonly, however, validation of cognitive tests aims to determine the acceptability of using test scores to infer behaviour on theoretical underlying constructs (Anastasi & Urbina, 1997; Delis, Jacobson, Bondi, Hamilton & Salmon, 2003; Murphy & Davidshofer, 2005; Foster & Cone, 1995; Hogan, 2003).

3.3 Construct Validity

Constructs are the theoretical attributes hypothesised to underpin test items and the phenomena upon which cognitive test development and interpretation are most frequently based. Construct validation is essential to the use of cognitive tests as it provides the primary empirical means by which battery structure may be determined. Cronbach and Meehl (1955 p.283), originally defined a construct as "some postulated attribute of people, assumed to be reflected in test performance", a definition which is apt to be repeated (i.e. Larrabee, 2003). Constructs are not defined in terms of observed performance on test items: rather test items are evaluated in terms of their capacity to reflect the latent construct underlying the test (Kane, 2001). Variation within test scores is assumed to indicate variation within the latent construct. Inferences based on this assumption rely on the validity of the test

as a measure of the intended construct, a test attribute which must be empirically determined (Foster & Cone, 1995; Nunnally, 1978).

As Foster and Cone (1995) state "much of psychology appears to accept constructs as the subject of study with little recognition of alternative" (p.249) and this is clearly the case in the field of cognitive assessment. Cognitive tests are commonly intended to tap variability in underlying cognitive constructs. Tests items constructed with this aim, as opposed to those which are intended to completely define the behaviour of interest, provide a sample of the test taker's underlying abilities, characteristics or skills in a particular domain of cognitive functioning. However such test items do not completely define the domain and instead are a sample drawn from a myriad of potential domain-specific items. If measurement using the items is free from the occluding effects of systematic error, then inferences drawn from the resulting test scores will provide meaningful or valid information about construct functioning.

While the methods by which constructs are empirically identified is the subject of some debate, the users of cognitive tests, like other scientists, hold firm beliefs about the "more prominent observables" which relate to the constructs with which they are concerned (Nunnally, 1978). In fact, the cognitive assessment literature reveals a marked consensus regarding meaningful domains of cognition. Reviews of clinical practice and test compendia alike indicate that test users regularly employ conceptions such as "memory", "visuospatial processing", and "attention" to describe the hypothesised cognitive traits or abilities theorised to underlie performance on cognitive test items.

Compendia, such as those produced by Lezak, Howieson and Loring (2004) and Strauss, Sherman and Spreen (2006) structure clinical thinking regarding

cognitive domains by presenting tests according to the constructs that they are hypothesised or demonstrated to measure. In the most recent edition of the seminal *Neuropsychological Assessment*, Lezak, Howieson and Loring, for example, structure tests according to those which measure orientation (i.e., attention and concentration), perception (i.e., verbal, auditory and tactile perception), memory, language, constructional abilities (i.e., drawing and assembly), reasoning (i.e., mathematical, visual or verbal reasoning), executive functions and motor performance.

Accordingly, practice reviews indicate that similar conceptions of cognitive domains guide test selection, analysis and interpretation in cognitive assessment. Butler, Retzlaff and Vanderploeg (1991) indicated that clinicians typically tested memory, speed of processing, language, visuospatial abilities, psychomotor abilities and executive functioning in addition to using most or all of the age appropriate Wechsler intelligence scale subtests (presumably to evaluate visuospatial, and verbal, if not attentional and speeded performance). More recently, Rabin, Barr and Burton (2005) reported substantial consensus regarding cognitive domains which clinicians routinely evaluated for diagnostic, descriptive and planning purposes. In this study a majority of clinicians "frequently evaluated" attention, verbal memory, executive function, visuospatial skills, nonverbal memory, intelligence, language and construction.

In many instances, however, the demarcations between "constructs" are uncertain and much overlap occurs in terms of the actual cognitive behaviours elicited by tests. Cognitive test items are likely to tap performance in several hypothesised constructs. Performance on a test which purportedly taps and is likely to be interpreted as a measure of memory, for example, may require verbal or visual

processing, spoken or graphomotor output as well as attentional abilities in addition to those which are purely memory based. The typical multi-dimensionality of cognitive tests presents substantial challenges to construct validation. In fact, authorities (Nunnally, 1978), recommend the combination of "a number of measures of such observables" to provide scope and sufficiency to measurement (p.98).

In psychometric terms, it would be intuitively appealing to consider the "true score" discussed in terms of psychometric reliability as synonymous with the "underlying construct" which forms the basis for many validation procedures. While it is arguably the aim of much test construction to achieve such a feat, it is rarely the case. The impact of occluding factors, including the influence of multiple constructs, is difficult to eliminate when such factors vary systematically. Most commonly, any given observed score will vary due to random and systematic error, and additional cognitive ability, as well as the underlying construct and both the reliability of the score and the validity of the proposed inferences are equally essential for clinical best-practice. Notwithstanding this unavoidable truth, it can be recognised that construct validation goes some way towards determining the meaningfulness of observed test scores (Franzen, 2000).

Nunnally (1978) prescribes "step-by-step" procedures by which construct validity can be established: first the entire domain of "observables" for a construct is identified; next these "observables" are evaluated (using empirical and statistical methods) in terms of their tendency to measure the construct of interest; and finally experimental measures of the construct are evaluated empirically in terms of adherence to construct based theory. Typically, Nunnally observes, construct validation occurs at the final level, that is at the level of the instrument, with previous steps being omitted.

The lack of systematic method in validation leads to a process in which constructs come to be defined largely by the combined evidence accrued from evaluation of the measures intended to tap them. As indicated above, stable cognitive constructs are indicated in testing and psychometric literatures most frequently through the use of factor analytic techniques (Delis, Jacobson, Bondi, Hamilton & Salmon, 2003; Henson & Roberts, 2006). In fact, shared variance techniques define the constructs used to produce many of the clinical conclusions drawn through the use of cognitive tests and while such techniques themselves are vulnerable to several problems, reviews of tests, journals, and the manuals of many commonly used cognitive measures reveal such techniques to be the most commonly used method of construct validation to date.

3.4 Validity and Clinical Decision Making

Construct validity is perhaps most profitably considered from the perspective of the battery rather than that of the individual test. In fact, clinicians are increasingly required to establish the scientific bases of battery-related conclusions (Bigler, 2007; Russell, 2007; Russell & Russell, 2003; Russell, Russell & Hill, 2005) providing pragmatic reasons for the more stringent application of psychometric validation to cognitive inferences drawn from batteries of tests. In fact, the need to apply validation in some form or another to cognitive test batteries has become increasingly necessary in light of criticisms regarding the scientific validity of assessment techniques (Reed, 1999). Demands to empirically establish the application of the scientific method to battery construction and use have arisen from a variety of sources including the legal profession (Bigler, 2007; Daubert v. Merrell, 1993; Reed, 1999; Reed, 1996; Russell, Russell & Hill, 2005; Russell & Russell,

2003), medical specialties (AAN, 1996), and managed care organisations (Ambrose, 1997). Questions regarding the validity of psychological assessments have, in some cases, led to criticisms from those outside the profession regarding instrument and battery type (Lezak, 2002). For example, Eisman, and colleagues (2000), in a review of over five hundred clinicians, cited increasing resistance by third-party payers in the United States of America to funding psychological assessments and restrictions relating to testing time and type of instrument deemed appropriate based on the empirical "validation" of the measures. More recently, Bigler (2007) recounted a "motion to exclude" flexible battery evidence based on an argument that this "approach had never been properly validated" and was "therefore, 'not reliable'" (p. 46). Fixed battery supporters, such as Russell, Russell and Hill (2005) have emphatically forwarded the belief that "fixed" batteries provide "dependable evidence" while, in contrast, flexible batteries cannot be considered to be valid as entire instruments. While challenges to the admissibility of cognitive battery results, based on empirical validation of the battery as a whole, have predominantly occurred in the North American legal system, what is deemed to constitute scientific "best practice" on one continent has little reason to differ on another.

Psychological ethical codes require that clinicians identify and use psychometrically reliable and valid measures. At the level of the battery, this requires empirical validation of the inferences which will be drawn from the battery as a whole. The fact that "validity" is the topic of some debate complicates efficient resolution of the dilemma and while fixed battery adherents are quick to extol the "validity" of their measures (Hom, 2003; Russell, 2007; Russell & Russell, 2003; Russell, Russell & Hill, 2005) the specific inferences typically validated for fixed batteries may not match the inferences required by clinicians.

For example, the HRNB aims to quantify the behavioural correlates of brain damage with the aim of neurological diagnosis (Reitan & Wolfson, 1985). To this aim, validation of the battery has explicitly focused on identifying the "aspects of psychological test results [that] relate specifically to the biological condition of the brain" (p.3). Validation of the HRNB has focused on the capacity of battery aggregate scores to accurately re-produce neurological diagnoses obtained medically (e.g., identify overall brain-damage, to lateralise and localise the anatomical location of brain-damage, and to differentiate between a "chronic, static lesion" versus a "recent, acutely destructive or rapidly progressive lesion"; Reitan & Wolfson, 1985, p.4). While the empiricism of this validation process cannot be faulted, and while the ever-changing nature of a flexible or semi-flexible battery does not facilitate research based evaluation of the capacity of a battery to differentiate between normal and brain-impaired individuals (Bigler, 2007), such inferences are arguably not the task for which the majority of cognitive assessment is conducted (Bigler, 2001; Lezak, 2002; Rabin, Barr & Burton, 2005). In fact, the most useful inferences may be those which provide a comprehensive description of an individual's cognitive functioning and behaviours based on specific cognitive domains, rather than any indices of "impairment". The true nature of the problem may, in fact, be a lack of method for formally applying validated cognitive constructs to flexibly constructed batteries.

Clinicians use cognitive domains which may be based on research literature, test manuals, availability, or habit to structure test selection, analysis and interpretation. The assumption that fixed cognitive domains underlie test performance is very widely accepted and is clear, for example, in the way that clinicians report the results of cognitive assessment. Donders (2001a, 2001b) for

example, reported that a majority of clinicians in a sample of 414 used cognitive domains to organise test findings in reports (54.35% “always”; 34.54% routinely” and only 2.42% “never”). It is not unreasonable to expect that clinicians organise the reporting of results according to how they have organised their test selection and combinations. Evaluation of functioning on these underlying or “latent” domains constitutes the real focus of cognitive assessment and provides the (implicit or explicit) structure for test grouping, analysis and interpretation.

To the extent that test combination is not formally based on validated domains, however, systematic error enters the resulting inferences. Cattell (1964) argued that to accurately choose from the large variety of available tests required of each clinician a “high degree of clarity and sophistication” in understanding the “universal parameters by which tests are evaluated” (p.1). This is equally applicable to current world practice, when the number of available measures has increased exponentially since Cattell’s observation in the 1960’s. Systematic error, which cannot be evaluated using CTT conceptions of reliability, but which still exerts an influence on the observed test score, can arguably be controlled at the battery level by better structuring the flexibly chosen battery according to empirically validated cognitive constructs which aid in the valid analysis and interpretation of cognitive test results (Haynes, Richard & Kubary, 1995).

3.4.1 Construct Validation at the Battery Level

The determination of empirically validated constructs relates directly to one of the most fundamental assumptions underpinning test usage (Anastasi & Urbina, 1997; Hayes, Richard & Kubary, 1995) which is the conception of cognitive test items as partial measures of the hypothetical constructs which are theorised to

underpin test behaviours. Variation in underlying constructs is inferred from variation in test items despite the fact that such items do not completely define the underlying constructs into which they tap (Foster & Cane, 1995). To compensate for the insufficiency of individual test items, they are combined into individual tests. In much the same way, individual tests can be combined into composite measurements of cognitive domains.

Construct-specific clinical inferences will benefit, in terms of increased validity, when they are based on combination of tests which measure the same construct. In fact several authorities recommend this both to strengthen validity and increase reliability. Nunnally (1978, p.98), for example, specifically argues that combination of several construct-specific measures facilitates clinical inferences about the "domain as a whole" attended by increased psychometric validity. In fact the combination of several construct related tests for the purposes of clinical inference is a methodology that has been employed in educational (Woodcock Johnston: Woodcock & Mather, 1989), cognitive (WISC-IV), neuropsychological (LNNB; Golden, Purisch & Hammeke, 1988) and personality (NEO-PI-R; Costa & McCrae, 1992) testing. These fields differ widely, however, in the degree to which such combinations are "formalised".

Many tests of unique, elementary cognitive tasks may be available to the clinician to measure functioning on a single unitary cognitive domain, the "construct" (Horn & Noll, 1994). In choosing tests for a semi-flexible cognitive battery, the clinician is in effect designing a unique, and un-validated, instrument from this plethora of choice. As Foster and Cone (1995) indicate, validation of one "category" or scale of a multi-scale instrument does not ensure the validity of the rest of the instrument. This is equally true of the cognitive battery, where the validity of

construct measurement overall is fundamentally limited by the validity of the least valid measure of that construct from which inferences are drawn.

In using batteries clinicians are likely to combine several measures of a single construct in interpretation. For example, clinicians may use the Information, Vocabulary and Similarities subtests from the WAIS-III in conjunction with the Boston Naming Test and Controlled Oral Word Association Test to determine "verbal functioning". While each of these measures may be subject to some form of empirical validation, the validity of inferences drawn from either formally or *informally* combined test scores is unknown. Even when tests are similarly valid measures of an underlying verbal construct, the relationship is unlikely to be acknowledged by "formal" test score combination, as is likely to occur in a fixed battery. Empirical evaluation of the degree to which each individual test contributes to measurement of the theorised underlying construct is necessary for valid inferences to be drawn regarding functioning on the construct overall. As validation pertains to the inference and not the test (Cronbach, 1971, Foster & Cone, 1995; Nunnally, 1978; Franzen, 2000), it logically follows that any inference required of a test, or combination of test scores, should be based on empirically evaluated meaning. Clear knowledge of the validity of the structure of cognitive functioning used is essential to determine test selection and combination which will lead to the most generalisable and accurate inferences.

Implicit to this argument, is the assumption that measurement of cognition is a measurement of unique domains and modalities. This is not always the case. The current conceptions of cognitive domains by which clinicians almost universally organise cognitive assessments (e.g., memory, attention, language, visual abilities, etc.) are theoretical in conception and in many cases defined by the very tests with

which they are purportedly measured. As Goldstein (1997) astutely noted, while brain functioning may be conceptualised as a specific system of domains it could just as logically be described by any useful and comprehensively constructed model. That is, any model of cognition which facilitates useful inferences could justifiably be used to structure a cognitive battery, as long as a systematic model is actually applied. Failure to use a validated model will, however, lead to several errors of both a psychometric and clinical nature.

3.4.2 Errors Associated with Lack of Valid Structure

The psychometric strength (reliability and validity in terms of sensitivity and specificity) and meaningfulness (clinical utility, predictive validity and ecological validity) of test interpretations is significantly compromised by a failure to apply a valid structure of cognitive domains to test selection and combination. Such errors relate largely to the increased likelihood of under- and over-evaluation of cognitive domains when non-systematic selection techniques are employed.

Under-evaluation of relevant cognitive domains compromises the clinician's ability to comprehensively and accurately evaluate all relevant cognitive behaviours. Sufficient tests must be given to ensure measurement of the domain is robust against, for example, the effects of anxiety or normal variability in test performance (Cimino, 2000; Russell & Russell, 2003). Failure to test domains sufficiently, may lead to the clinician missing evidence of impairment (or skill) through failure to elicit the relevant behaviours (Russell & Russell, 2003).

Conversely, over-evaluation of domains leads to an increased chance of finding impaired performances simply due to chance (Ingraham & Aiken, 1996).

Additionally, over-emphasis on a given domain may make it seem more important, biasing interpretation of test results (Wedding & Faust, 1989).

Finally, when highly correlated tests are used to measure a domain they are likely to support each other. This means little, however, in terms of interpretation, as such tests measure the same or highly similar aspects of the underlying behaviour and thus add very little unique information to the clinician's understanding of domain functioning. Failing to evaluate the possibility of multicollinearity could lead to biases in interpretation of test results.

Hawkins and Hastie (1990) suggest that the integration of "large, interdependent, implication-rich, ambiguity- and contradiction-bearing evidence", such as that provided by the test scores from a typical cognitive battery, must begin with some means of simplifying and ordering the data (p.323). Valid constructs may be used to impose this order on a semi-flexible cognitive battery. Most commonly such constructs are established through the use of factor analytic techniques.

3.5 Factor Analysis

"Factor analysis is intimately involved with questions of validity. . . [and] is at the heart of the measurement of psychological constructs" (Nunnally, 1978, pp.112-113). Factor analysis, including exploratory factor analysis, confirmatory factor analysis, and principle components analysis is fundamentally a data reduction technique which facilitates the grouping of a large number of scores based on measures of inter-correlation (Tabachnick & Fidell, 2000; Thompson, 2004). It aims to do so without a significant loss of information (Gorsuch, 2003). Factor analytic techniques are highly useful to test users who also seek to simplify interpretation of a large number of individual test scores by grouping those theorised to measure the

same underlying "construct" together. In basic terms, factor analysis acts to "explain a larger set of j measured variables with a smaller set of k latent constructs" (Henson & Roberts, 2006, p.394). When any given factor analysis produces factors which explain a majority of the variance caused by associations between the original variables, these groupings become highly meaningful and useful to the test user. Specifically, such factors are frequently believed to be the cause of observed scores (Gorsuch, 2003; Henson & Roberts, 2006; Kieffer, 1999; Thompson & Daniel, 1996). In fact, validity theory has grown in close conjunction with shared variance techniques and from conception has shared similar, if not mutual, aims (Cronbach & Meehl, 1955; Guilford, 1948; Kaplan & Saccuzzo, 2005). A fundamental aim of construct validation is the identification of theorised psychological traits which are indirectly measured by psychological tests and which are conceptually synonymous with the "factors" empirically produced using shared variance procedures (Larrabee, 2003).

Several authors cite the pivotal role of factor analytic techniques in the establishment of "constructs" which form the interpretative basis of much clinical examination. Cronbach and Meehl (1955) in their seminal discussion of construct validity listed factor analysis as one of the several means by which construct validity may be established. Correlation based procedures remain by far the most commonly employed means by which cognitive constructs are evaluated and strong arguments exist for the judicious use of factor analytic techniques in establishing construct validity (Jacobson, Delis, Hamilton, Bondi & Salmon, 2004; Larrabee, 2003; Tulsky & Price, 2003). Delis, Jacobson, Bondi, Hamilton & Salmon (2003), for example, cite over sixty studies presented at the annual meeting of the International Neuropsychological Society, 2002, which used factor analytic techniques or inter-

variable correlations to identify or define cognitive constructs. Given this pivotal role, it is not surprising that use of factor analysis and related procedures is the subject of some debate in the psychometric and testing literature.

3.5.1 Caveats to Use of Factor Analysis

As Franzen states (2000, p. 27), evaluation of validity in terms of "how we know what we know" cannot be separated from method, that is "techniques for evaluating the validity of instruments" and this is fundamentally true of factor analysis and construct validity. Current understandings of cognitive structure have been developed via what may be termed, somewhat irreverently, a process of "psychological bootstrapping". That is, theoreticians such as Thurstone, Wechsler, or Halstead, hypothesised about cognitive functions and devised tests to examine these hypotheses. Analysis of the results of these tests was used to define cognitive functioning: the definitions of which were fundamentally limited by the scope of the original theories. The fact underlying this is that statistical techniques, regardless of how robust or sophisticated, are fundamentally limited by the data upon which they are employed. Hence, use of methods such as factor analysis to define a structure of cognitive functioning is vulnerable to the variables which enter the analysis, which in turn, are limited by the scope of existing cognitive tests. It is increasingly clear that any conceptualisation of cognitive domains derived solely from statistical analysis of cognitive test data can be "comprehensive" only to the degree that the test data entering analyses is "comprehensive". As Henson and Roberts (2006) and others (i.e. Mulaik, 1987; Thompson & Daniel, 1996) note the meaningfulness of constructs depends fundamentally on "researcher definition".

In other words, if a theorist wishes to understand cognitive domains employing factor analytic techniques and the subtests of the WAIS-III, a finite number of results can be expected (e.g. some variation of the familiar verbal, visual, attentional and speeded indices; Wechsler, 1997a). The understanding of cognitive domains gained from such an analysis must be limited by the finite number of possible relationships between the thirteen WAIS-III subtests. While long term memory or executive functioning are widely considered as legitimate domains of cognitive functioning (Lezak, Howieson & Loring, 2004; Mitrushina, Boone & D'Elia, 1999), they cannot possibly be produced by such an analysis as the scope of content of these factors is lacking. The principles or theories of cognitive structures do not drive the testing rather the tests available drive the principles. Stated emphatically, cognitive domains are determined by data which is gathered from cognitive tests which fundamentally biases the establishment of a statistically pure understanding of cognitive constructs, highlighting the need for what Henson and Roberts (2006) refer to as "thoughtful" researcher judgement when validation is conducted.

Perhaps this is less of a dilemma when the nature of constructs themselves is considered. As Nunnally (1978) rather bluntly comments, rather than the objective and empirically sound entities which would be implied by a "foolproof" mathematical procedure, the constructs produced as the result of validation analyses are in effect "something that scientists put together from their own imaginations" (p.96). If the means with which cognitive constructs are identified is imperfect, this is reflective of the entity being studied. While this is in no way an excuse for the cavalier use of shared variance techniques, it does highlight the situation-specific

nature of constructs, which is demonstrated clearly in the psychometric and testing literature.

For example, current debate has centred on the efficacy of factor analytic techniques, employed on large normative and clinical samples, to evaluate construct validity given the likelihood that the interpretable constructs produced by analyses will vary based on sample composition. Delis, Jacobson, Bondi, Hamilton & Salmon (2003), conducted factor analyses of immediate, delayed-recall and recognition measures using both normative and "mixed" clinical samples comprised of subjects suffering either Alzheimer's disease or Huntington's disease or both. Results of these analyses indicated that memory processes, which formed a unitary factor in the normative sample, dissociated in the presence of relevant brain pathology. Jacobson, Delis and colleagues (2004) further clarify their criticism to pertain to the use of heterogeneous samples for such investigations. According to this research group, construct validation required a "systematic, programmatic exploration involving separate confirmatory factor analyses using multiple homogenous patient populations" (p.1020). This conclusion is further highlighted by factor analyses conducted by Jones, Schaik and Witts (2006) on a sample characterised not by clinical impairment, but by low IQ scores. These authors were unable to extract a robust four factor solution, and instead indicated a two factor solution analogous to the traditional verbal and performance split as best fitting the data. While many theoreticians would disagree, this highlights the necessary subtlety with which factor analytic techniques must be employed in construct validation particularly when the samples differ from "normal".

This point is again well demonstrated using the theoretical "best-case scenario" of thorough and repeated validation using factor analytic techniques, the

WAIS-III and WMS-III. Despite conscientious validation these batteries have a domain structure strongly influenced by this fundamental limitation and are liable to substantial changes based on sample composition and the variables included in analyses, amongst other things. Analyses of the WAIS-III and WMS-III standardisation sample indicates a four factor solution for the WAIS-III (i.e., verbal processing, visual processing, attention and concentration and speeded process) and a five factor solution for the WMS-III (i.e. immediate, delayed, visual, verbal and working memory; Wechsler, 1997a). Similar solutions have been supported by subsequent published research (i.e. Arnau & Thompson, 2000; Ryan & Paola, 2001; Saklofske, Hildebrand, & Gorsuch, 2000). Factor analytic study of the *combined* WAIS-III and WMS-III standardisation sample by Tulsky and Price (2003), however, indicated a six factor structure including verbal, perceptual, processing speed, working memory, auditory memory and visual memory constructs. This deviates from the structures produced by analyses of the WAIS-III and WMS-III samples separately and has held true for other normative (Bowden, Castairs, & Shores, 1999) and clinical samples (Bowden et. al., 2001; Bowden, Book, Bardenhagen, Shores & Castairs, 2002).

Additionally Tulsky and colleagues (2003) report that subtest scores for WMS-III memory tests dissociate into visual memory, verbal memory and working memory without an immediate and delayed dissociation in “normal” samples. These authors contend that adhering to this structure in the evaluation of “cognitive compromised” individuals would rob the clinician of highly salient dissociations between immediate and delayed recall likely to occur in the clinical population (Tulsky, Ivnik, Price & Wilkins, 2003). Such results clearly demonstrate caveats to

the application of statistically derived factors to providing concrete information about cognitive domains in a clinical setting.

Bowden (2004), further highlighted several issues of relevance to the use of shared variance techniques in construct validation which may produce similar variation in interpretable constructs. Specifically, this author cautions against the impact of unrepresentative sampling and changes in score reliability, or variability between groups. Bowden argued, however, that such cautions "in no way lessen the value of the factor-analytic approach" (p.1018). The need for a greater depth of understanding in interpretation of factor analytic results does not provide sufficient argument against its use in the validation of cognitive constructs.

Use of factor analytic techniques as a means of identifying robust and definitive domains is vulnerable to deviations from the "reality" of the hypothesised cognitive constructs, based on the strength of the mathematical procedure itself, the degree to which the tests used as variables actually estimate variation in the underlying construct, and to the intrinsically undefinable nature of the constructs themselves. More succinctly, factor analytic representation of cognition is vulnerable to deviations from "reality" based on the degree to which the sample deviates from "reality". Factor analytic techniques will provide no definitive answers, but must instead be used as aids to the decision making process. If deviations in dissociable factors between samples and studies are inevitable, the astute researcher may make use of this artefact as meaningful in itself. In this way, factor analysis is conceptualised as a powerful tool in the complex process of accurately operationalising human cognitive behaviours, rather than a definitive authority.

3.5.2 Use of Factor Analysis to Structure the Battery

Construct validity specifically provides an empirical structure for test selection, reducing battery usage errors which relate to misattribution of inferences based on domains and other error judgements typical to flexible battery construction. Therefore, despite these caveats, and occasionally because of them, the use of the cognitive constructs indicated in factor analytic research to structure a semi-flexible battery is still likely to result in a reduction in decision making errors in several instances.

1) Validity is Fundamentally Related to Reliability: An empirically validated structure is necessary to capitalise on the psychometric strengths provided to clinical decisions by evaluation of reliability. Specifically, factor analytic techniques used in construct validation indicate dimensionality and are highly relevant to the correct application and interpretation of internal consistency coefficients, as discussed in chapter two. As Cortina (1993) indicates the vulnerability of coefficient alpha as a measure of reliability for multidimensional tests necessitates the use of factor-analytic techniques to ensure minimal multi-dimensionality or to at least alert test users to its presence and allow for the appropriate interpretation of reliability coefficients.

2) Factors Tend to be Robust and Give Meaning to Deviations: While factor or component structure is likely to change based on variables included and sample composition, factors tend to be robust within normal samples. The appearance of robust factors in such samples may not specify the structure for specific clinical settings, but does give strength to interpretations. When discrepancies are found between tests, which load strongly on the same factor in a “normal” sample, weight is added to a hypothesis that performance is impaired. For example, the Boston

Naming Test loads closely with WAIS-III Vocabulary as a test of verbal ability in normal individuals. Dissociation between these two tests, such as is likely in a patient with nominal dysphasia for example, strongly indicates the presence of a specific disorder.

3) "Factor jumping" actually demonstrates the fundamentally multi-modal nature of many cognitive tests: This characteristic is well recognised by clinicians. For example, Rabin, Barr and Burton (2005) cite that of the 273 memory tests, 220 attention tests and 219 tests of executive functioning reported by clinicians, only 101, 45 and 56 respectively were unique to their domains: the remaining tests were used interchangeably in clinical interpretation. Factor analytic research may in fact provide a rationale for this clinical practice (Kane, 1991) and at least highlights the delicacy with which some cognitive tests must be interpreted.

Specifically, the presence of impairment on multi-modal tests cannot be automatically attributed to impairment on the underlying domain and instead may be due to impairments in input channels (e.g., visual, auditory or tactile), processing requirements of the tests (e.g., phonological, orthographic or semantic processing; Craik & Lockhart, 1972; Jacoby, 1983) or output channels (e.g., verbal, written, pointing or naming responses; Bauer, 2000). Impairment in any of these modalities may result in poor test scores even without any compromise of domain-specific abilities. For example, while the WAIS-III subtest Arithmetic is intended as an attentional measure and indeed loads on the Working Memory factor, an impaired score may be attributable to deafness, dyscalculia, depression or distraction, as well as "working memory". Similarly, a test of visual-spatial skills, such as WAIS-III Block Design, may require working memory, motor skills and visual-spatial perception. As Franzen eloquently states, "although it may be more cumbersome to

describe the test as assessing visual-spatial, motor-reproductive skills for which short-term memory is required, it is less cumbersome than the theoretical excess baggage required to explain discrepancies in performance between a situation that requires memory and a situation that does not require memory (p.30)". The impact of these often meaningful factors must be considered when attributing the cause of impaired test performance despite even well established construct validity (Cimino, 2000).

To accommodate multidimensional tests, clinicians may choose to structure assessment to better test the influence on test performance of factors other than these "traditional" cognitive domains (Bauer, 2000). For example, inconsistent patterns of impairment on more well accepted cognitive domains may be resolved by structuring test scores into domains based upon specific modalities of input (such as aurally presented information, written information, symbolic information or sensory information), processing (semantic, symbolic, educationally based) or output (spoken, written or drawn responses, pointing responses or naming responses). An alternate, modally developed structure may reveal specific deficits which would be hidden by the structure of cognitive domains commonly prescribed by factor analytic research on normal populations. However, in accomplishing this task clinicians will be impeded by a lack of existing analytical structure with which to interpret alternative combinations. All of this could require the clinician to closely study the relevant factor analyses, or given sufficient available data, conduct factor analyses of samples specifically relevant to their own unique client base.

When cognitive test results are interpreted clinically it is useful to believe that cognitive constructs have a "true nature" which exists regardless of theory or measurement methods (Ruscio & Ruscio, 2002). Psychometric measurement aims to

approximate these "true" constructs with the least degree of error. As Ruscio and Ruscio (2002) indicate the latent structure of psychological variables remains relatively unexplored despite the accumulated wealth of research (Foster & Cone, 1995). In the absence of appropriate empirical investigation, however, a construct's true latent structure cannot be presumed to match the tangible scale with which it is sampled. In other words, the onus is on the clinician not to exceed the interpretations which are justified empirically when behavioural inferences are conducted using results from both individual tests and batteries.

Further, and more importantly, while a wide range of factor analytic research may be employed in the structure and interpretation of a battery it must be used with due consideration of the vulnerabilities and strengths of the technique in describing valid cognitive constructs for the individual client. Kane (1992) commented that modern assessors still have far to go towards identifying the fundamental abilities which underlie complex, cognitive functions. Given the caveats discussed above, factor analytic research could be used in a much more sophisticated manner in the structure of cognitive batteries. Despite the complexity, however, the search for valid domains is unavoidable as clinicians require formalised cognitive structure.

3.6 Applying Factor Analytic Research to Battery Structure and Interpretation

Factor analytic techniques undoubtedly provide valuable suggestions for battery structure, however, cautions must be exercised when using empirical constructs to select and combine tests in a semi-flexible battery. In fact, factor analysis may provide mild, moderate or strong influences on battery structure.

At the lowest level of impact, factor analytic research indicates that tests may measure the same construct or disparate constructs and such research is used to

suggest test combinations. For example, tests which load on the same factor across a variety of studies may be logically considered to measure, at least in part, the same underlying construct. These tests may be recognised for their similarity and drawing inference from them in combination should provide a more comprehensive measure of the construct and one which to some degree ameliorates the potentially conflicting influences of the various other domains which impact on individual measures. On the other hand, tests which tend to "jump" factors between different studies should be used with more caution given empirical evidence of their fundamentally multi-modal natures.

At a more precise level, the factors which emerge from a particular analysis may be used to define the constructs upon which test selection and combination are based. Scores from the tests which load on these constructs or factors may then be combined to provide a composite measure of the underlying construct, with the likely advantage of increased reliability. While there are strong psychometric arguments for structuring a battery specifically around the results of factor analytic research, however, constraints, such as the availability of a suitable analysis, may impinge.

At the most precise level, factor loadings may indicate the weighting of each test of a particular construct. Tests for a given factor would then influence decision making regarding the underlying construct according to the precise degree to which they provide a pure measure of that construct. For example, a test which has a factor loading of .5 on a particular factor would contribute approximately half that of a test which loads 1 on the factor. This technique has been used with success in personality test research using the MMPI-2 in assessing personal-injury claimants (Goh, 2006).

This degree of application is warranted when the individual being assessed is effectively represented by the sample from which factors are derived. However, while this degree of precision seems optimal, it could arguably lead to an unwarranted degree of confidence in the validity of construct measurement. In other words, the results of factor analyses vary according to the sample participants and according to the specific included variables hence factor loadings from one sample cannot be considered precisely applicable to individuals not from that sample. Use of factor loadings to weight a composite factor score would therefore produce a misplaced degree of precision unless the variables were the same as those included in the factor analysis and the individual was highly similar to those in the original sample.

More pragmatically, weighting of a composite measure according to factor loadings is likely to prove prohibitively difficult as test selection would be limited to those measures included in the factor analysis. This is very possible in certain fixed measures (e.g., WAIS-III, WMS-III) but would be nearly impossible to approximate for a flexibly changing test battery. It would be unfortunate for clinicians to miss altogether the benefits to be obtained via consideration of factor analysis by the setting of overly stringent requirements. As the dilemma facing clinicians, is how to accommodate the technical data into actual clinical practice, Table 3.1 outlines recommendations for improving the application of validity research to the structuring of a semi-flexible cognitive battery. As indicated in the table, varying degrees of application of factor analytic research are warranted, largely based on the degree to which the individual measured is represented by such research.

Table 3.1

Levels at which Factor Analytic Research May Be Applied

	Type of Information	Application
Level 5	Robust factor structure obtained from clinical setting specific to the client, using the same battery of tests.	Factor loadings can be used to calculate weighted composites of tests scores for each factor if all tests in factor analysis are administered. Otherwise, factor loadings indicate logical test combinations based on cognitive domains.
Level 4	Robust factor structure obtained from very similar clinical settings.	Consistent factor loadings indicate tests likely to measure the same construct and which should logically be combined in analysis and interpretation.
Level 3	Robust factor structure obtained from normative data.	Factor loadings indicate tests likely to measure the same construct in an unimpaired sample. These can be combined, however, the clinical meaningfulness of combination should be critically evaluated.
Level 2	Less robust factor structure obtained from clinical or normative settings.	Consistent factor loadings indicate tests measuring each construct and should be used to structure test combination. Care should be taken when adding obviously multi-modal tests to composites.
Level 1	Domain structure obtained from theory of test similarity.	Unless a theoretical similarity is backed up by factor analytic research, assuming the construct validity of the resulting test combinations may introduce error into inference.

3.7 Conclusions

Validity relates to the second fundamental consideration of assessment clinicians: that of clearly establishing the meaning of clinical inferences. Validity is perhaps of foremost relevance in the difficult task of drawing meaningful conclusions about the cognitive functioning of individuals using tests. Drawing such inferences without due reference to validity is likely to introduce error into measurement.

The process of validation is less amenable to mathematical description than, for example, reliability and perhaps for this reason has been the topic of much debate. Specifically experts disagree regarding the most appropriate validation processes and a wide range of methodologies are available to suit a myriad of potential clinical inferences. In cognitive assessment, however, validation is most likely to pertain to the structure of cognitive domains as clinicians most frequently comment on domain-specific functioning.

Information regarding cognitive domains is largely gathered through factor analytic methods. While such techniques are vulnerable to errors of sampling, technique and interpretation, consistent understandings of validated cognitive structures are available in the psychometric literature. However, the clinician would do well to apply factor analytic research to the structuring of his or her cognitive battery with more specific regard to the suitability of the sample from which factors are derived. In fact, the process of validation is an ongoing one, in which the clinician may be actively involved and those who administer cognitive tests in real world environments are arguably best placed to empirically evaluate the degree to which those tests may be used to draw useful conclusions about clients (Wilkinson & Robertson, 2006).

Valid cognitive structure is essential to battery structure. As is often the case in clinical practice, however, precise ordering of the battery based on psychometric principles is impeded by various practical requirements. To state this frankly, while test "X" of a given construct may be the most reliable measure available and have undergone considerable validation, if it is in English and the test taker speaks only Mandarin it is clearly not the correct choice. In fact, practical constructs legitimately dictate battery structure and analysis more frequently than psychometric considerations, as discussed in the following chapter.

CHAPTER FOUR

STUCTURING A CLINICALLY FLEXIBLE BATTERY

4.1 Introduction

Test selection based solely on psychometric considerations may still fail to provide a comprehensive, accurate or highly meaningful description of relevant behaviours. Psychometric theory aside, each clinician must eventually choose from the finite number of tests at his or her disposal, a specific collection of measures to be used in eliciting the most relevant cognitive behaviours for the individual client. Individuals present with differing demographic characteristics (e.g. gender, culture, age, educational level) and degrees and causes of impairment (e.g. educational difficulties, head injury, vascular disease or dementing disorders), and may require assessment for a wide variety of purposes (e.g. rehabilitation planning, prediction of future recovery, degree of compensation). Such differences in assessments are likely to be reflected in construction of the assessment battery (Bauer, 2000; Ogden, 1996). In fact, clinicians must integrate psychometric considerations with a variety of complex practical constraints which are the third fundamental concern of clinicians in compiling and using groups of cognitive tests.

Clinicians tend to rely on a small common group of tests. Rabin, Barr & Burton (2005) cited that while clinicians occasionally chose from over 270 tests for the evaluation of memory, they most commonly utilised just three measures: 71% in this sample used a Wechsler memory battery (WMS/WMS-R/WMS-III); 54% used the California Verbal Learning Test; and 45% utilised the Rey-Osterrieth Complex Figure Test. While test selection and organisation may be based on well-validated cognitive constructs, rarely can or should factor analytic research form the sole rationale for test selection and interpretation. Similarly, while the choice of reliable

measures seems paramount in the psychometric literature, even the most reliable tools may be rejected if they do not facilitate relevant clinical inference. Instead, the cognitive domains and relevant tests to achieve optimal clinical meaningfulness are frequently dictated by characteristics of the test setting, the clinician and the individual client which perhaps explains why fixed batteries fail to be universally adopted in clinical practice, despite the various strengths associated with their potential for formalised structure and actuarial interpretative methodologies.

Accommodating the frequently competing demands of client and test setting, however, presents a fundamental challenge to the psychometric soundness of the battery. Despite this, clinicians modify practice according to the needs of the individual client. For example, Donders (2001a, 2001b) found that clinicians modified the organisation and content of reports in response to the clinical presentation of the client and specific knowledge regarding relevant base-rates, group appropriate norms and uses for reported data. Rabin, Barr and Burton (2005) found clinicians reported use of abbreviated and more focused instruments and suggested this could be in response to increasing time limitations for assessment. Some consideration of practical, clinical factors is an inevitable aspect of using a semi-flexible cognitive battery.

Practical considerations, in fact, provide pressing reasons why client and setting factors must be accommodated in cognitive assessment. For example, failure to consider relevant variables increases the likelihood of false attribution of variance to error, which should rightly be attributed to the systematic influence of the disregarded variables (Franzen, 2000). Such errors occur when cognitive evaluators ignore the impact of evaluation context, individual demographics, learning history,

low motivation and affective states on test performance. Clearly, no amount of psychometric robustness can compensate for poor test selection.

If cognitive assessment focuses exclusively on statistical and psychometric measurement of intellectual behaviours, understanding of the individual, which requires the clinician to gain understanding of the emotionality (feelings and motivations) and control (how behaviour is expressed) aspects of functioning as well, will never be achieved with any degree of accuracy. Misunderstanding of individual cognitive functioning is highly likely to occur if the systematic effects of non-test related factors are not, at least, considered, in addition to psychometric concerns. Such factors may be divided into "client factors" (i.e., age, education, gender, culture, disability), "test setting factors" (i.e., time, test availability, experience of the clinician) and "interpretive factors" (i.e., establishing equivalence of norms, need to evaluate multiple measures for each domain, need to determine how many tests measure a domain sufficiently, evaluation of hypotheses). Each of these may confound the quantification of cognition (Cimino, 2000; Lezak, 1983).

At this point, clinicians may lack the techniques to empirically evaluate the influence of such variables on test scores and must attempt to do so using a limited number of corrected norms, standardised test administration and sound clinical judgement. Test-inferences, however, can never be "valid" or "reliable" in a practical sense without some attempt to determine the impact of confounding factors. The difficulty of this task does not excuse the reticence to undertake it: "The price paid for . . . omission ranges from unfortunate (in the form of lowered validity coefficients), to inexcusable (in the form of misleading information resulting in disservice to the patient)" (Franzen, 2000, p.31).

In conducting cognitive assessment, the clinician must determine the scope of cognitive abilities to be examined, appropriate tests to examine these abilities, the depth of investigation and how test scores equate to actual brain-behaviour relationships (Bauer, 2000). Clinical practice reviews indicate that, regardless of the esoteric debates, while the pool of tests from which the battery is constructed may be selected based on psychometric principles the individual battery is just as likely to be developed according to clinical issues. The accommodation of such constraints, within the bounds of psychometric validity and reliability, presents the third primary challenge to assessment clinicians.

4.2 Client and Setting Factors

Even though a test score validly measures a particular construct, it may be influenced by several factors relating to the client and test setting. Such factors must be considered in test selection and include the culture, age, education, gender, mood, psychiatric history, drug use history, and likelihood for distractions in the test setting. While such issues are the most frequently considered in battery construction, they further complicate the task of compiling a practical and psychometrically sound testing battery and thus merit further consideration. In many ways the following discussion provides evidence for the pressing need for flexibility in battery construction and the related need for flexibly applied methodologies to evaluate the psychometric consequences of these inevitable battery modifications.

4.2.1 Practice Effects

Repeated testings of individuals are common in clinical, medico-legal or forensic assessment. However, practice effects may exert a substantial impact on the

resulting test scores by allowing clients to change their test-taking strategies (Basso, Bornstein & Lang, 1999), apply less attention and concentration (Lezak, 1995; Lezak, Howieson & Loring, 2004) and remember answers from previous assessments. The effects of practice on cognitive test scores are well documented in the literature. Basso, Carona, Lowery and Axelrod (2002), for example, demonstrated considerable increases in WAIS-III composite and index scores after both a three and a six month interval in a sample of normal participants.

If a pre-injury assessment is available, tests may be chosen to best facilitate the valuable comparisons between current and pre-morbid functioning. On the other hand, clinicians may need to avoid specific tests to prevent confounding due to practice effects from a recent previous administration. In both instances a degree of clinical flexibility in test selection and combination is clearly required which must also be addressed with knowledge of relevant validity information to guide the selection of alternative measures.

4.2.2 Impairment in Input or Output Modalities

Physical limitations of the client, such as physical disability, visual impairment or hearing impairment, may preclude use of certain tests and pose substantial challenges to battery construction and test usage (Bertone, Bittinelli & Faubert, 2007; Hill-Briggs, Dial, Morere & Joyce, 2007). To provide an obvious example, the visuospatial abilities of a vision impaired client may clearly not be tested using WAIS-III Block Design, which is ordinarily a very commonly employed visual test. This is particularly salient given that the client populations of many assessment clinicians may experience sensory or physical disabilities at significantly greater rates than the general population (Iezzoni, McCarthy, Davis, & Siebens,

2001; Park, Mayer, Moghimi, Park & Deremeik, 2005).

Hills-Briggs, Dial, Morere and Joyce (2007) outline comprehensively the impact of sensory or physical impairment on test selection, administration, modification and interpretation, and specifically discuss the fairness, accuracy and validity of modifying test administration or battery selection to accommodate impairments in input or output modality. These authors highlight the complexity and sensitivity of modifications that may be required to avoid biased inferences cautioning, however, that use of non-traditional test administrations or response formats reduces validity and may invalidate interpretive guidelines. Valid and reliable test usage requires of the clinician a thorough understanding of the impact of these factors on cognitive test performance, further increasing the complexity of the clinician's task in compiling a semi-flexible cognitive battery.

4.2.3 Culture

Rarely is the need for clear-headed and informed tailoring of a multi-measure test battery to the needs of the individual better illustrated than by reviewing the impact of culture on the performance of cognitive tests (Manly & Echemendia, 2007). Culture, as it refers to the differing race or ethnicity, and resulting values, attitudes, interpersonal behavioural styles and communication conventions unique to certain groups of individuals, impacts on the performance of cognitive tests in systematic ways. These are frequently related to access to formal "Westernised" education, literacy, acculturation and socio-economic status (Ardila, 2005; Baird, Ford & Podell, 2007; Brauer-Boone, Victor, Wen, Razani & Ponton, 2007). Systematic errors in testing caused by factors of the examiner, test setting or test administration must be largely controlled through standardised assessment which,

however, may fail to generalise to non-Western cultures characterised by very different first language and conventions regarding authority, communication processes, knowledge, and time. Specifically, differing cultural exposure to timed testing procedures (Agranovick & Puente, 2007), culturally related lack of access to formal education (Nirini et al., 2004; Ostrosky-Sol´ıs, Ramirez, & Ardila, 2004; Reis, Guerreiro, & Petersson, 2003), degree of acculturation (Kennepohl, Shore, Nabors & Hanks, 2004), and culture-related socio-economic and psychosocial status (Byrd, Touradji, Tang & Manly, 2004; Manly, Jacobs, Touradjui, Small & Stern, 2002; Manly, Touradjui, Tang & Stern, 2003; Ryan, Baird, Mindt, Byrd, Monzones & Morgellow, 2005) have all been found to systematically and significantly impact upon the performance of cognitive tests.

Cognitive assessments are “based on the assumptions as well as the values of scientific and technologically oriented societies" (Ardila, 2005, p. 186). Some authors contend that the very concept of cognitive assessment is a concept largely specific to Western culture. This may substantially complicate the task of compiling a practically appropriate testing battery, which is both psychometrically and clinically useful, for an individual client not of a Caucasian, North-American cultural background (i.e., for which cultural group a large majority of test norms have been developed). This issue is particularly salient when cognitive tests are translated, in terms of language, into various culturally-specific “official forms” (Gierl & ElAtia, 2007). Such translation in fact has the potential to mask issues of culture.

4.2.4 Education

Many cognitive assessments have been developed to evaluate individuals who have undertaken some level of formal education and cognitive tests frequently

rely on similar tasks to those encountered during school. The impact of education on the performance of cognitive tests is expected given the fundamentally educational nature of many of the original cognitive tests. This was recognized early in the development of psychometric cognitive testing (Wechsler, 1935; 1939) and discrepancies in literacy and access to formal "education" have been convincingly linked to differences in neuropsychological and cognitive test performance (Manly, 2005). The influence of formal education has been demonstrated in cognitive tasks such as naming (Ardila, 2007; Baird, Ford & Podell, 2007; Kennepohl, Shore, Nabors & Hanks, 2004; Lezak, Howieson & Loring, 2004), pattern matching (Baird, Ford & Podell, 2007; Byrd, Jacobs, Hilton, Stern & Manly, 2005), visuospatial ability, verbal ability, speeded performance and attention (Wechsler, 1997a) and some memory tasks (Folia & Kosmidis, 2003).

As years of education increase, performance on many cognitive tests can be expected to increase. This general trend occurs because such testing specifically taps skills and abilities gained during the course of formal education and common to those individuals who developed the majority of their cognitive skills in this particular setting (Ardila, 2005). The cognitive processes necessary to the acquisition of formal education, the particular memory styles and the specific abilities, such as reading, writing and mathematical manipulation, gained during the course of schooling can all be expected to impact on the performance of cognitive tests, are implicit in the formation of many normative groups, and hence must be considered in the interpretation of cognitive test scores.

In addition, errors in diagnosis of impairment are frequently caused by failure to take into account the age and premorbid ability levels of the client (Matarazzo, 1990; Russell & Russell, 2003). Russell and Russell (2003) cite experienced

clinicians making diagnostic mistakes based on such glaring errors as applying children's norms to adult test takers and then concluding "normal" performance despite the fact that such a normative comparison would fail to provide adequate proof for any conclusions about the adult client. It is, however, well accepted clinical practice to apply age and grade adjusted norms to data interpretation (Donders, 2001a; 2001b), which goes some way towards remedying potential interpretive errors. Again, however, this provides a strong argument for clinical flexibility in battery construction which cannot be met by a fixed collection of cognitive tests.

4.2.6 Specific Learning Disability

Observed scores may be impaired for a wide variety of reasons, unrelated to the underlying construct of interest. Frequently, this may be related to specific educational disabilities which can profoundly impact upon the performance of cognitive tests. As discussed previously, while the WAIS-III Arithmetic subtest loads on a Working Memory factor, in the presence of a specific learning disability in mathematics, the subtest is undermined as a measure of attentional capacity. Accommodating this phenomenon in battery construction may require the clinician to integrate validity research with a process of hypothesis testing, as client disabilities become apparent.

4.2.7 Specific Clinical Conditions

On a related note, tests that load on a particular construct in factor analytic studies of normal or mixed clinical samples may be differentially vulnerable in specific clinical groups. Such tests may include those indicative of specific

educational deficits, such as dyslexia or dyscalculia, or specific neuropsychological or cognitive impairments, such as aphasia. It should be noted that much factor analytic evaluation of tests is conducted on "normal" samples the cognitive behaviours of which are clearly markedly different to those of a clinical population (Heaton, Taylor & Manley, 2003). In order to avoid the impact of tests which may be specifically vulnerable to different cognitive states, the clinician must be aware of factor analytic research. This was thoroughly discussed in the preceding chapter but serves to highlight here, again, the need for clinical flexibility in battery construction.

4.2.8 Test Setting

Cognitive tests are administered in many disparate testing environments, not all of which suit the "optimal" testing conditions. In order to minimise errors due to fatigue, time constraints, third party observers, distractions, noise, or, interruption the clinician must choose efficient means of eliciting the behaviours of most relevance to the required inferences. Tests may also be chosen for their dual purpose. For example, WRAT-3 or -4 Reading may be administered as a verbal (word knowledge) and educational (reading level) test, which is an efficient means of evaluating two constructs with a single administration. Flexible batteries may be chosen for their brevity whereas a major barrier to the use of fixed batteries in clinical practice is their length. For example, the HRNB is estimated to take four hours to administer by an "experienced" examiner and includes a large battery of tests (Reitan & Wolfson, 1985).

Tests may also be chosen for the flexibility with which they may be analysed to accommodate for unavoidable deviations from standardised procedure. Again, many fixed batteries necessitate, by definition, unchanging administration and

analysis and any deviation from standardised procedures invalidates pattern analysis techniques which rely on known and unchanging relationships between tests (Golden, Purisch & Hammeke, 1988; Reitan & Wolfson, 2004).

In compiling and analysing a cognitive battery, the clinician must often deviate from habitual practice to accommodate various factors of the client and test setting. Modern clinical practice frequently necessitates a high level of flexibility in test selection and combination. While clinicians routinely accommodate clinical considerations, the integration of such logistical concerns with a robust level of psychometric strength is difficult to achieve at the battery level using currently available methodologies. In light of discussion in the preceding chapters, however, the need for a methodology to achieve this is pressing.

4.3 Normative Factors

In addition to consideration of clinical factors, clinicians must consider several interpretive factors which may also necessitate flexibility in test selection and combination. This is frequently related to the suitability of particular normative samples.

Typically, cognitive assessment makes use of comparisons between the performance of an individual and the combined performances of a group of normally functioning individuals. In conducting a cognitive assessment norms are frequently used both descriptively to indicate an individual's level of cognitive functioning in relation to a reference group, and diagnostically as evidence of cognitive impairment; (Busch, Chelune & Suchy, 2005; Heaton, Taylor & Manly, 2003; Mitrushina, Boone & D'Elia, 2005). Clinicians may need to consider available normative samples in terms of their specific composition, their similarity to an individual client, and their

suitability for comparison with normative samples of other tests in the battery. Again, the accommodation of these factors within a psychometrically valid and reliable cognitive battery must be considered in construction and analysis.

4.3.1 Evaluation of Test Normative Sample.

Norm based inferences may be compromised by vulnerabilities of the normative sample and norming techniques used. Hiscock (2007) summarised analytical errors occurring due to: normative data which inaccurately represented the intended population due to inadequate sample size or the collection of data using a method which is not replicated in clinical use; normative data which was confounded due to systematic demographic differences from the population of use; and, normative data which was "old" or outdated. Each of these potential vulnerabilities could impact upon the usefulness of normative based inferences.

The representativeness of a normative group is fundamentally reliant on the sampling of adequate numbers from the population of interest (Crawford & Howell, 1998). Typical normative comparisons require use of the mean and standard deviation of the normative group to standardise the observed score through a linear transformation which may then be compared to the normal curve to determine significant deviation from the mean and rarity within the sample which is assumed to be normally distributed. Alternately, frequencies within the normative sample may be used to approximate the normal curve and produce look-up tables which associate a particular raw score with a scaled or standard score indicative of certain position on the normal curve, a procedure known as uniform norming. Both methodologies are sensitive to sampling inadequacy, uniform norming being perhaps slightly more so (Howell, 1997) and the meaningfulness of normative comparisons clearly depends on

the adequacy of the entire normative sample, or sub-groups of the sample based on relevant demographic factors (Hiscock, 2007). Some authorities have suggested methods to accommodate inadequate sampling by using alternative comparison techniques such as considering the individual as a sample of N=1 and conducting a modified t-test to determine significant deviation from the mean (Crawford & Howell, 1998; Sokal & Rohlf, 1995). However, few techniques can accommodate for inadequate normative samples.

Currency of norms is another fundamental challenge to drawing useful and meaningful norm-based inferences. Clearly re-norming must accompany updates in test content or modifications in administration or scoring procedures (Hiscock, 2007). Most importantly, however, updated norms are needed to accommodate for systematic increases in mean cognitive test performances which have been demonstrated over time across cultures (Flynn, 1984; 1987; 1994; 1998a, 1998b, 1999).

4.3.2 Suitability of the Normative Sample to the Individual Client

Much of observed score interpretation relies upon normative comparisons. However, the efficacy of normative comparison is useful only to the extent that a comparison between the individual test taker and the normative sample is warranted by demographic similarities. Systematic errors in inferences may be due to systematic discrepancies between the client and the normative sample and observed deviations from the "norm" are only meaningful when the observed individual can be expected not to deviate if normal functioning is preserved. Russell, Russell and Hill (2005) state somewhat testily that "no neuropsychological sets of norms have come close to representing the whole average population of the United States, with the

exception of the Wechsler tests" (p. 791). This norming inadequacy in the United States is clearly exacerbated when tests are used in geographic areas where little or no culturally appropriate norming has occurred.

Discrepancies between the demographic characteristics of the test taker and the normative samples increase the likelihood that normative comparisons lack meaning or are entirely inappropriate. In other words, normative comparisons increasingly lack meaning as the individual deviates from legitimate membership of the population from which the normative group is sampled. As discussed above, such membership is dependant on a variety of factors which may all impact upon performance of cognitive tests, including gender (Gale, Baxter, Connor, Herring & Comer, 2007; Herlitz, Airaksinen & Nordstom, 1999), age (Whittle et. al., 2007), education (Wechsler, 1997a), culture, ethnicity or language of subjects (Echemendia, 2007; Gollan, Fennema-Notestine, Montoya & Jernigan, 2007; Gollan & Fennema-Notestine, 2007), or membership in a specific generation (Flynn, 1999; Neisser, 1998). While not all of these factors are likely to influence performance on all cognitive tests, their potential relevance to the individual client should be evaluated in the course of test selection. If, for example, age and education are shown to produce systematic variation within test scores then normative comparison should be conducted based on relevant age and educational categories.

If the effects of demographic variables are successfully controlled, the confounding effects of such variables are modified. This is particularly necessary if test scores from variant normative samples are compared during analysis. In compiling a cognitive battery, the clinician must consider the appropriateness of the normative samples against which observed score comparisons will take place.

4.3.3 Suitability of Normative Comparison Between Tests

Finally, interpretation of a battery of cognitive tests may require comparison between two or more measures within the battery. Resulting discrepancies may, however, be attributable to disparities between normative samples due to sample composition, consideration of demographic factors, currency, scaling and raw score distribution (Kalechstein, van Gorp & Rapport, 1998). Establishing the equivalence of normative samples and of the scale units of test scores within the battery is essential as these provide two primary sources of systematic variation within test scores which cannot be attributed to variation within the underlying construct of interest (Anastasi & Urbina, 1997).

Specific challenges are presented to clinicians by discrepancies between normative raw score distributions. Cognitive tests have two primary scaling types. Achievement scales, in which score elevation coincides with increased ability are common in tests such as the WAIS-III and WMS-III. For example, in these batteries a scaled score (SS) of 19 on a subtest indicates superior performance; SS of 10 is "average"; a SS of 2 indicates "extremely low" performance (Wechsler, 1997a). Such measures have most often been developed with the measurement of "intelligence" in mind. On the other hand, tests developed from a neuropsychological perspective are often impairment scales, on which elevated scores indicate a greater degree of impairment. Such scales are present in tests such as the Trail Making Test (Reitan & Wolfson, 1985) and provide specific challenges for comparisons between test scores in a battery.

An underlying and more complicated normative disparity occurs due to fundamentally different normative distributions between the two scaling types. Impairment scales typically occur on tests designed to target pathology in impaired

individuals. These tests tend to be "easy" and the low ceiling is likely to be rapidly reached by "normal" individuals. Impairment scales fail to provide subtle gradation between performance in the upper and "average" ranges of performance, while specifically defining various degrees of poor performance (Ardila, 1999). The normative sample of such a test will be heavily skewed, with the majority of the sample falling in one end of the distribution. Conversely achievement scales may have ceilings which are unhelpfully high for use with clinical populations and hence fail to specifically describe the degrees of lowered performance, while accurately defining average to above average ability on the tests. Scale type is therefore indicative of potential underlying normative disparity which could make combined interpretation of such tests difficult even when they are validly considered as measures of the same construct.

A less subtle source of error is that due to disparities between scaling units. Disparities between scaling units are simply corrected, however, by standardisation of individual scores into z-scores from which linear transformation to any common metric (i.e., any score with a specified mean and standard deviation) is possible (Anastasi & Urbina, 1997; Lezak, Howieson & Loring, 2004; Mitrushina, Boone & D'Elia, 1999; Nunnally, 1978). Comparison between standardised test scores may then be conducted with accuracy, a methodology which is employed in fixed batteries, such as the WAIS-III and WMS-III and which may be adapted for use with other collections of tests.

One of the most compelling characteristics of the fixed battery is the potential for developing co-norms for the entire group of tests alleviating potential discrepancies between the normative samples of individual tests and facilitating test scale unit compatibility. Unfortunately, as discussed above, this advantage is

entirely compromised if such normative samples do not match with individual clients regardless of comprehensive, co-normed construction. For example, despite the established impact of culture on performance of cognitive tests, no fixed battery has been convincingly normed on other than a Northern American population (Ardila, 2005). Additionally, regardless of the recognised impact of currency on normative based inferences (Hiscock, 2007) norms for the primary fixed batteries are also universally outdated.

Authorities concur that establishment of normative equivalence on various levels is essential to prevent error in multi-modal battery interpretation (Anastasi & Urbina, 1997; Ghiselli, 1981; Nunnally, 1978). This may be done through both mathematic and logical means. For example, the equivalence of test scores can be evaluated by a comparison of their respective percentiles within each normative group. Scores which share a common percentile rank may be considered equivalent, a process termed the equipercentile method (Anastasi & Urbina, 1997). Alternately, the development of normative data specifically suitable to each test setting could be undertaken by practicing clinicians. While such "local" norms are highly appropriate to individuals of the target population, compilation of sufficient numbers to control for sampling and measurement error is a lengthy process. Regardless, the choice of tests which may be viably interpreted in terms of normative comparisons presents a final challenge to the clinician in compiling a cognitive battery which facilitates meaningful inferences about functioning in the individual. This process is doubly complex when domain-based comparisons are required.

4.4 Considering Clinical Factors in Test Selection and Combination

The changing aims of cognitive evaluation motivate discussion of these practical concerns to cognitive battery usage. As Bialystok and Craik (2007) somewhat acerbically comment, early cognitive theory and norming relied largely on participants who were, “18–21 years old, right handed, largely male. . . enrolled in Introductory Psychology classes at major universities . . . English speaking, middle class and reasonably intelligent” (p.209). On a fundamental level this tendency has changed as clinicians and researchers seek to understand the cognitive functioning of a much broader scope of individuals.

Similarly, neuropsychological test batteries were originally developed to diagnose the presence of impairment (Halstead, 1947), however, due to the rapidly increasing precision of neuroimaging techniques diagnostic efficiency is rarely the aim of a test battery today (Bauer, 2000; Bennett, 2001; Bigler, 2001; Lezak, Howieson & Loring, 2004). Neuropsychological and cognitive tests cannot be used to determine the presence and anatomical locus of brain damage with anything approximating the accuracy and reliability provided by neuroimaging techniques, such as magnetic resonance imaging, functional magnetic resonance imagining, computed tomography, and positron emission tomography (Arai, 2005; Binder, Frost, Hammeke, Cox, Rao & Prieto, 1997; Demonet & Thierry, 2001; Dougherty, Rauch & Fischman, 2004; Fernandez-Duque & Posner, 2001; Frisoni et al., 2003; Goldstein & Price, 2004; Kaneko, Momose & Kadoya, 2005; Leaper et al., 2001; Liotti & Mayberg, 2001; Park & Gonzalez, 2004; Savoy & Gollub, 2004; Yancey & Phelps, 2001). For this reason, test batteries are increasingly used to "understand the nature of organic deficits" and hence are developed for their "usefulness in eliciting different kinds of behaviours that are relevant to the patient’s condition and needs"

rather than for predictive accuracy (Lezak, 1995, p. 686). To achieve adequate behavioural description a multi-modal battery must sample behaviours from the "major cognitive functions across at least auditory and visual, verbal and non-symbolic modalities (Bauer, 2000; Lezak, Howieson & Loring, 2004). Reviews of the clinical literature indicate that test selection for a battery must be moderated by the degree that administration and use is suitable, practical, useful and interpretable.

From a clinical perspective, “construct validity” and the appropriate levels of test and composite reliability must be decided in conjunction with a consideration of which tests, constructs and test combinations provide the most “meaning”. During this process clinicians must consider aspects of the client, test and interpretive methodology, in addition to psychometric characteristics of the battery. Battery usage in cognitive testing necessitates the consideration of practical factors as well as psychometric concerns, which are summarised in Table 4.1.

As summarised in the table, the accommodation of practical concerns can be complex and in some instances clinicians will experience difficulty in deriving meaningful normative interpretations. In fact, consideration of the psychometric qualities of unavoidable test combinations provides clinicians with essential information regarding the confidence with which they can interpret subsequent scores.

Table 4.1

Clinical and Practical Considerations in the Structure of a Semi-Flexible Cognitive Battery

	Type of Information	Application
Level 5	Client has full vision, hearing and mobility, average education, no specific learning disabilities and is a North American of Caucasian ethnicity. Test performance is not influenced by motivation, emotion/mood or inability to complete testing. Tests are co-normed	Subtests are readily comparable, based on established equivalence of norms. Available norms are likely to be highly suitable for the individual client. Normative comparisons may be undertaken with a high level of confidence, provided subtest reliability is suitable. Test performance is less likely to be confounded by input/output deficits.
Level 4	Client has a single impairment in sensory, physical, educational or motivational capacities OR ethnicity other than North American Caucasian, but similar to (i.e., British, Australian, New Zealander) OR subtests are normed with strongly normal raw score distributions.	Subtests must be chosen to avoid confounds due to impairment. Norms are likely to be equivalent and comparison requires linear transformation of subtest scores. Norms are likely to be suitable for the individual client. Normative comparisons may be undertaken with moderate confidence, provided subtest reliability is suitable.

Table 4.1 (continued)

	Type of Information	Application
Level 3	Client has several impairments of sensory, physical, educational or motivational capacities OR ethnicity substantially different to North American Caucasian OR some subtests with raw score distributions deviating from normality.	Subtests must be chosen to avoid confounds due to impairment. Norms may not be equivalent; consider alternate tests. Norms may be unrepresentative of the client. Normative comparisons must be undertaken with caution.
Level 2	Client has several impairments of sensory, physical, educational or motivational capacities AND ethnicity other than North American Caucasian AND/OR Primary language other than English OR many subtests with raw score distributions deviating from normality.	Subtests must be chosen to avoid confounds due to impairment. Norms may not be equivalent; consider alternate tests. It is highly unlikely that norms are representative of the client; normative comparisons must be undertaken with caution.
Level 1	Client has substantial impairments in sensory, physical, educational or motivational capacities AND ethnicity substantially different to North American Caucasian AND Primary language other than English AND/OR non-normal raw score distributions	Test scores are likely to be confounded by language, impairment or culture. Normative comparisons are difficult to justify with a strong suspicion that norms are very unrepresentative.

4.5 Conclusions

The task of clinicians in compiling cognitive batteries is a complex one in which the competing demands of psychometric theory must be integrated with potentially more pressing clinical necessities. Implicit in this is the necessity of due regard to practice effects, specific impairment modalities, client culture, client levels of education, specific learning disabilities and other client clinical conditions, and various factors of the test setting, all of which could impact upon the accurate measurement of cognitive abilities. In addition, normative factors such as the robustness of normative samples, the suitability of available norms for the individual client and the appropriateness of normative comparisons given lack of co-norming between flexible battery measures must be considered when tests are selected and analysed. These sometimes complex requirements must be considered to avoid the occlusion of cognitive measurement by artifactual errors.

In fact, integration of psychometric and clinical constraints is achievable if the appropriate methodology is used. Readily available actuarial techniques for score combination, based on direct reference to factor analytic research and appropriate usage of available reliability methods may facilitate both the psychometric strength and clinical robustness of battery based inferences. Such a methodology would allow for necessary flexibility in battery construction, while providing necessary information about the psychometric consequences of clinical decisions. This is discussed in the following chapter.

CHAPTER FIVE

USING CLINICAL COMPOSITES

5.1 Introduction

The previous chapters aimed to highlight the magnitude of the task facing clinicians in attempting to "make sense" of the principles of psychometric theory in practical terms. These chapters discussed the on-going challenge these processes present to clinicians and theoreticians alike and concluded that current practices are unlikely to completely control for the myriad of potential errors intrinsic to the task of assessment. Given the complexity of the preceding theoretical discussions, a brief summary of how these fundamental issues apply to clinical decision-making using test battery scores is warranted.

When using cognitive batteries, the clinician must consider reliability, validity and clinical utility. When cognitive tests are used clinically, assessors must choose, administer, score, analyse and interpret test scores with varying degrees of interrelatedness and psychometric robustness and in response to a wide range of referral questions. Consideration of psychometric and clinical requirements is therefore perhaps best undertaken in light of their impact during test selection, analysis and interpretation. Each of these stages of assessment requires specific consideration of psychometric theory and clinical usefulness. In fact, the issues of psychometric reliability, validity and clinical utility (specifically normative issues) share an interactive relationship and may be considered at each of the three stages of battery development and use. The pervasive influence of theoretical considerations on practical clinical tasks necessitates consistent attention when cognitive tests are used in combination, as outlined in Table 5.1.

Table 5.1

Relationships between Psychometric and Clinical Tasks of Battery Usage

	Reliability	**Validity**	**Normative Issues**
Test Selection	Individual test reliabilities, based on samples appropriate to the client, must be used to select stable measures.	Validity of tests for measuring relevant cognitive domains must be used to select measures to elicit relevant behaviours.	Usefulness and appropriateness of available norms must be evaluated to ensure subsequent normative comparisons are meaningful.
Test Analysis	Reliability of test combinations must be considered to ensure stability and indices of errors calculated, providing a means of directly considering random error.	Validity of test combinations must be considered to ensure that meaningful inferences may be drawn using test scores.	Normative sampling and demographics must be considered to allow application of appropriate normative adjustments.
Test Interpretation	Reliability of domain based inferences must be considered to indicate the confidence warranted by the impact of error on scores.	Validity of domain-based inferences must be applied to moderate the confidence with which domain-based inferences may be drawn.	Functioning on specific domains must be examined in relation to performance of normative samples. Functioning must be examined to indicate discrepancies between domains.

At the first level, reliability, validity and standardisation issues apply to the task of test selection. Considerations of test reliability dictates both the type and number of tests selected, factor analytic research indicates measures empirically established to validly tap relevant cognitive constructs, while the representativeness of available norms indicates the likely meaningfulness of subsequent normative comparisons. Most importantly at this level, the overall battery may be structured according to psychometric theory to avoid various errors of clinical judgement to which an unstructured collection of test scores is particularly vulnerable. At this level of battery usage, the number and type of tests are chosen to produce the most stable, accurate and meaningful conclusions about relevant cognitive domains within the constraints of the client. Despite pressing psychometric and normative demands, pragmatic factors such as the practicality of the test for clinical use and availability of and familiarity with the test are likely to exert more compelling influences on the types of tests included in the battery. This carries the advantage of improving the capacity of the battery to accommodate unique characteristics of the individual client. In this circumstance, however, consideration of psychometric issues becomes perhaps more pressing. In fact, achieving battery structure which is both theoretically and practically sound relies on knowledge of psychometric validity, the internal consistency of measures and various clinical considerations without which inferential errors are highly likely.

The analysis of test scores, that is the transformation from raw scores to interpretable indices, such as a scaled score, standard score or percentile rank, is subsequently influenced by theoretical considerations. Reliability facilitates the direct application of error indices to clinical decision making and could add valuable information to domain-based analyses. Validity indicates tests which may be

analysed collectively with the aim of inferring functioning about relevant cognitive constructs. In addition, factors of the normative sample, such as scaling, distribution, various methods of standardisation, and the potential need for demographic adjustments guide normative analyses

Ultimately cognitive assessment is conducted with the aim of drawing accurate conclusions about the cognitive functioning of individuals. This process is likely to be structured within a framework of practically meaningful cognitive domains, such as “verbal functioning”, “attentional ability”, “processing speed” or “memory”. Despite their appealing familiarity, the underlying assumptions regarding the relationships between test scores and these constructs are strictly psychometric. As with preceding stages, ignoring psychometric considerations at the test interpretation stage prevents clinicians from evaluating the impact of error on this process. Errors which occur during battery construction and analysis make erroneous inferences unavoidable. Even when clinicians successfully reduce or control for errors at the structural and analytical stages of battery usage, the need to address measurement error is on-going. At the interpretative level, the reliability of interpretations provides valuable information regarding the impact of error on the battery. The use of an empirically validated structure during preceding levels facilitates domain-based interpretations. When interpreting test scores clinicians may need to establish the functioning of the individual in relation to a normative group (normative) or comparisons between an individual’s functioning on various domains (ipsative) may be required. Needless to say, errors arise when such test scores indicate different levels of functioning or when systematic methods of establishing the significance or abnormality of discrepancies, which take into account varying reliability, are not used.

It is clear that clinicians wishing to be more fully informed about the myriad of potential error sources in assessment must undertake a matrix of complex considerations throughout the stages of assessment. It is not surprising that attempts to integrate the large collections of ambiguous data points which typify cognitive assessment are impeded by a degree of justifiable confusion. In fact, it is likely that clinical inferences are drawn about cognitive domains and comparisons made between individuals' functioning on cognitive domains without reference to differing levels of reliability and interrelationship, variations within scaling and normative samples, and varying validity for the required inference.

Despite the complexity of the task, issues of usefulness, accuracy and meaning should serve an appropriately fundamental role in the compilation, analysis and interpretation of cognitive test scores, a task to which clinicians and theoreticians alike should aspire. In response to this, a reliable approach to psychological testing (RAPT) is proposed which aims to integrate psychometric and clinical concerns. There are two distinct yet related components to the RAPT methodology: theory-based guidelines for the selection, structure, analysis and interpretation of a semi-flexible battery of cognitive tests; and the tools and procedures required to generate and analyse composite scores. This methodology ultimately aims to address the complexity facing the majority of clinicians who modify their testing battery to accommodate practical constraints.

5.2 RAPT Methodology

The first aspect of the RAPT methodology is the configuration of a semi-flexible battery of test scores into domain-based composite scores, based on empirically validated constructs and guided by principles of reliability and clinical

utility. This methodology operationalises the discussions of reliability, validity and clinical factors conducted in the preceding chapters. In RAPT it is proposed that domain-based composite scores form the basis for test score combination, analysis and interpretation due to the resulting increase in reliability and consequent reduction in measurement error associated with composite-based inferences. This approach exemplifies a conceptual process which may be used to guide cognitive test usage from selection through to inferences regarding cognitive functioning.

The second component of the RAPT methodology is the algebraic formulae which facilitate combination of test scores into domain-based composite measures. Most of these formulae are readily available in the psychometric literature and several are routinely incorporated into existing interpretive frameworks, such as the Wechsler scales. Application of these formulae into an integrated analytical framework for a semi-flexible cognitive battery as proposed in this thesis, however, has not occurred in the testing literature to date.

The first aspect of the RAPT methodology is conceptual in nature and suggests specific guidelines to formalise the incorporation of both psychometric and practical considerations. The second aspect of the RAPT methodology is computational in nature and provides a "toolbox" which may be used in analysing and interpreting the results of cognitive assessment according to psychometric guidelines. In combination, these aspects of RAPT methodology may be applied to the use of cognitive tests at various levels with varying degrees of impact, ranging from the generation of modified composite scores within existing battery frameworks to better evaluate clinical hypotheses, to the development of entire structures of battery usage based on psychometric theory and adapted to specific clinical conditions. Current clinical assessment practices seem lacking in explicit methods

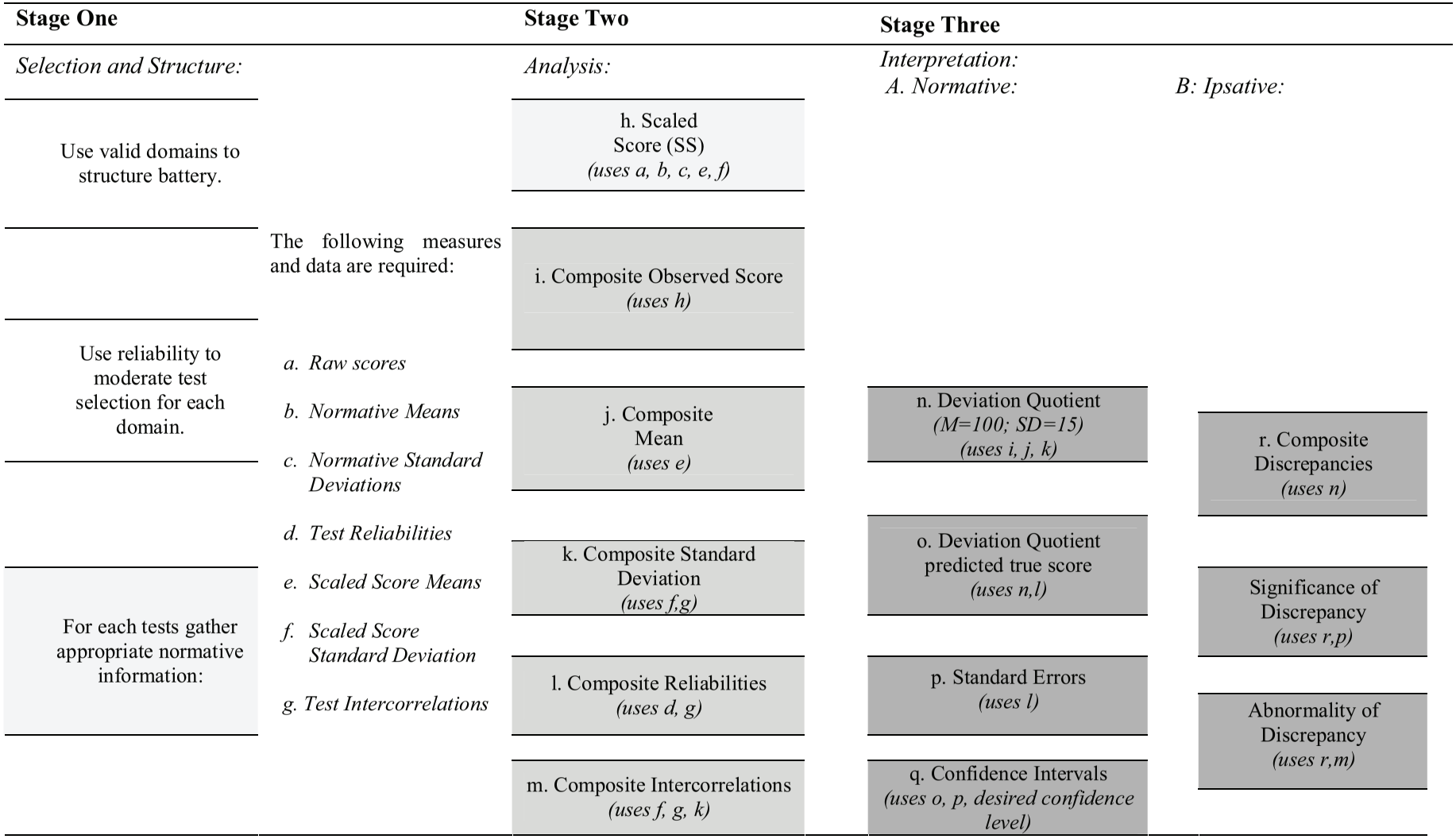
that produce acceptably accurate, dependable and meaningful results. It is hoped that elucidation of the RAPT methodology will constitute a practical contribution to the production of more stable, accurate and useful results from cognitive assessments.

5.2.1 RAPT Model

RAPT methodology applies psychometric theory to the three main stages of battery construction, usage, and, analysis and interpretation, as outlined in Figure 5.1. As indicated in this figure, stage one is the selection of a semi-flexible test battery based on clearly validated cognitive constructs. Tests are selected based on their reliability and on their normative and practical suitability to the individual client. The reliability associated with the overall measurement of cognitive domains dictates the selection of an appropriate number of tests to measure each construct. During this stage the clinician collects relevant data for each test including normative means and standard deviations, and reliability coefficients. In addition, the intercorrelations between tests in the battery are collected. Tests are then administered and raw scores obtained and standardised into scaled scores (i.e., M = 10; SD = 3).

The rationale of using scaled scores at this level should be explained. Adoption of a consistent scale for standardised scores facilitates the combination of test scores into composites. While this could be achieved as easily using z-scores with a mean of 0 and a standard deviation of 1, the convention within testing has been to convert z-scores into one of a number of scaling systems. Use of the scaled score with a mean of 10 and a standard deviation of 3, although arbitrary, is widespread and its use should facilitate both recognition and uptake by clinicians.

Figure 5.1 RAPT Methodology



At this stage, however, it is important to note that conventions such as scaled scores ranging from 1 to 19 (as per the Wechsler scales) is unjustifiable, as is their representation as whole numbers. These conventions place artificial constraints on the range of standardised scores, imposing unnecessary ceiling and floor effects and resulting in a resolution which takes a continuous raw score distribution and represents it as a discontinuous scaled score distribution. For example in converting a raw score of 10 into a scaled score of 8 and a raw score of 11 to a scaled score of 9 results in a discontinuity between the 25th percentile (SS = 8) and the 37th percentile (SS = 9) with no representation of scores in between. This results, of course, from the entirely arbitrary practice of representing scaled scores as whole numbers. It is recommended that in the RAPT methodology scaled scores are represented at least to one decimal place to retain the continuous aspects of the raw score distribution.

The next stage (stage two) involves the generation of domain-related composite scores based on:

a) The observed scores of the component tests selected to measure each cognitive domain. Scaled scores of each domain are summed to form an additive composite score.
b) Computation of composite means through summing of the population means for the component tests. If scaled scores have been employed as a common metric for tests then the composite mean simplifies to the number of component tests multiplied by 10.
c) Computation of composite standard deviations using component test standard deviations and intercorrelations.
d) Computation of composite reliability using component test reliabilities and intercorrelations.

e) Computation of the intercorrelations between composites using the composite and test standard deviations, as well as the intercorrelations between tests.

The final analytical stage (stage three), includes both normative (Part A) and ipsative (Part B) analyses and interpretation of composite scores. For ease of interpretation, the composite observed scores are standardised into deviation quotients (DQ) represented as standard scores (M=100: SD=10) which may then be evaluated in terms of the standard normal distribution (i.e., using percentiles based on area under the normal curve).

As with the selection of a scaling system for tests, the use of the standard score at this level also merits explanation. Perhaps the most familiar deviation quotients in clinical practice are the IQ scores associated with many intelligence batteries. Their adoption here reflects their common usage in clinical practice for representing composites associated with domain-related constructs. However, they also serve a much more pragmatic purpose which is to facilitate and maintain the distinction between scores represented at the test level and those at the composite level. Use of these different scaling systems reduces the likelihood that practitioners could inadvertently mistake one score for another. This is reduced by the fact that the scores represented by these two scaling systems are unlikely to overlap in their practical ranges. Of course, these scaling systems are nothing more than linear transformations of the same underlying normal distribution (expressed as z-scores) but it can be readily appreciated that while mathematically simpler, representation of all scores as z-scores would be vulnerable to confusion and reduce the likelihood of detecting computational errors. As an additional observation, psychologists seem to be strongly averse to numerically representing any human behaviour with negative numbers.

The composite reliability coefficients are used to calculate predicted true scores (DQ_{PT}) to accommodate for regression to the mean around which confidence intervals are symmetrically placed using the relevant indices of standard error (SE_M, SE_E, SE_P). This accommodates for the likely impact of error on overall measurement of the domain. Finally, the overall profile of composite scores is evaluated for significantly or abnormally discrepant performances. In this step, the relevant discrepancies between composite DQ's are calculated and evaluated in terms of significance (i.e., taking into account the standard error of measurement) and abnormality (i.e., taking into account the composite standard deviations and intercorrelations). In combination, these inferential steps allow the clinician to answer several pertinent questions within an estimated degree of confidence.

5.2.2 Advantages of RAPT Methodology

Several advantages, in terms of psychometric strengths, may be obtained through use of RAPT methodology. First, test selection and interpretation according to an empirically validated structure of cognitive functioning potentially presents a challenge to clinicians. Using RAPT methodology, the battery is explicitly structured according to the cognitive domains suggested by factor analytic and other relevant research. More importantly, however, this empirical structure is consolidated by analyses in which tests of each domain are formally combined into composite scores, according to a rationale of clinical utility, and with a known degree of internal consistency and empirically established validity. In other words, the empirically established battery structure dictated by factor analytic research is formalised by the use of composite methodology.

Clinicians may currently lack an empirical rationale for the numbers of tests of each domain which are chosen, instead relying on such factors as test availability

or habit. However, successfully choosing an appropriate number of tests to measure each cognitive domain both thoroughly and without redundancy is vitally important to the reliability of resulting inferences. Using RAPT methodology the internal consistency of domain measurement using any collection of domain-specific tests may be evaluated, providing an empirical rationale for test selection. When the reliability of domain measurement can be calculated, domain specific tests may be added to the battery until the overall reliability of domain measurement ceases to improve. Subsequent tests may be included for clinical utility, however, the reliability of domain measurement will always be known allowing the clinician a degree of control over errors relating to over- or under-testing. RAPT methodology provides a method by which clinicians can be informed by the clinical literature and operationalised in the selection of tests for a cognitive battery with calculable consequences for reliability. Similarly, where clinicians are required through necessity or pragmatics to employ less than optimal measures, the impact of these choices upon reliability can be computed and accommodated in test analysis and interpretation.

Use of composite methodology specifically reduces the impact of impaired test scores due simply to chance or to factors other than the construct of interest. Using RAPT methodology sufficient tests of each cognitive domain are administered to achieve an optimal level of composite reliability without the over-evaluation of domains likely to result in a higher probability of artifactually poor test scores. Composite methodology reduces the impact of individual poor scores as the reliability of composite scores improves over the reliability of any individual test scores. The likelihood of a type I error, in which cognitive impairment is inferred when there is none, is reduced by use of the composite methodology proposed in RAPT.

The ubiquitous norms-based representations of cognitive assessments may also be improved by use of RAPT methodology. As discussed above, clinicians are often required to draw conclusions about overall functioning in cognitive domains such as verbal ability, memory or executive functioning. Again, the composite methodology proposed in RAPT allows direct empirical evaluation of the error associated with drawing such inferences through the use of indices of error in the subsequent construction of confidence intervals. The application of RAPT techniques provides clinicians with empirically established levels of confidence based directly on the reliability with which domains are measured.

Finally, the RAPT methodology facilitates evaluation of the significance and abnormality of discrepancies between cognitive composites. Clinicians are required to draw conclusions regarding deviations between the domain-based functioning of their clients. When compared with performance on other domains, relatively "impaired" performance on a collection of domain specific tests (i.e., visuospatial, verbal, or memory) may be indicative of specific disorder or at least of interest to the clinician. RAPT methodology allows such conclusions to be drawn using statistical techniques which take into account the error associated with measurement and the interrelationships (intercorrelations) between tests and domains. In this way clinicians can determine whether composites differ statistically from one another and estimate the frequency with which differences between composites occur in the general population based upon the normative data and the normal distribution.

Overall, RAPT methodology provides empirical answers to the challenges against reliability, validity and utility which are so frequently levelled at practicing clinicians. Using this methodology the clinician has concrete solutions to the fundamental questions regarding the reliability associated with domain-based test scores, the validity with which domain-based inferences may drawn and the

confidence with which conclusions can be postulated about the cognitive functioning of individuals. In addition, the relationship between the literature and data that informs clinical practice can be clearly linked to clinical interpretations and decision-making which is at the core of evidence-based practice. The utility of this approach is summarised in Table 5.2 which lists ten fundamental clinical questions and the relevant sections in previous chapters which support the RAPT methodological steps suggested.

Table 5.2

Clinical Questions Answered by use of RAPT Methodology

Clinical Question	RAPT Technique	Section
"What tests should I use in my battery?"	Test selection is based on structure indicated by valid cognitive domains and by factor analytically driven test combinations. Evaluation of the reliability associated with such combinations provides additional evidence of the psychometric robustness of choice.	• *3.5.2* • *2.5*
"How many tests should I include for each domain?"	Combinations of tests can be examined for their ability to reliably measure a specific domain. Where test selection is limited by test availability the consequences of these decisions can be determined and accommodated with regard to composite scores	• *2.5*
"What about artifactual error or poor test scores which are due simply to chance or factors other than the construct of interest?"	Test selection according to composite reliability prevents over- or under-evaluation of domains. Combining domain-based tests reduces the impact of any individual test scores, reducing type I errors.	• *3.4.2* • *3.4.1*

Table 5.2 (continued)

Clinical Question	RAPT Technique	Section
"How is this individual's verbal/visual/memory/etc compared to normals?"	Computation of percentile ranks and confidence intervals for all composites permits normative-based descriptions of performance with appropriate representation of measurement error.	• *2.3.2* • *4.3.3*
"Has this individual's verbal/visual/memory/etc changed in relation to a prior assessment?"	Computation of retest confidence intervals using standard error of prediction (SE_P) facilitates comparisons driven by statistical methodology.	• *2.3.2*
"Are there significant strengths or weaknesses across this individual's different cognitive domains?"	Significance of profile discrepancies is evaluated through tests of significance applied to difference scores between pairs of composites using standard error of measurement.	• *2.3.2*

Table 5.2 (continued)

Clinical Question	RAPT Technique	Section
"Are there abnormal strengths or weaknesses across this individual's different cognitive domains?"	Abnormality of profile discrepancies evaluated based on calculated inter-relationships between tests and composites (i.e., intercorrelations between tests and composites).	• *4.3.3*
"What is the reliability associated with my inference for each cognitive domain?"	Internal consistency of domain-based composites is directly calculated. The clinician may interpret scores for each domain in combination with a known degree of measurement error.	• *2.5*
"What is the validity associated with my inference for each cognitive domain?"	Inferences are directly related to the empirically validated domains on which composites are based. Composite methodology provides a means of formalising test score combinations with the advantage of increased overall reliability.	• *3.6*
"How confident am I in my assessment of this individual's cognitive functioning?"	Inferences are based on test scores with clearly evaluated psychometric properties and test scores are interpreted with known confidence levels. RAPT methodology additionally provides the flexibility to avoid reduced confidence due to clinical constraints (i.e., invalidated test scores, test setting distractions).	• *2.3* • *4.2*

5.2.3. Applying RAPT Methodology to the Fixed versus Flexible Debate

As discussed previously, the semi-flexible battery has been criticised strongly by fixed battery proponents due to its vulnerability to measurement errors. In fact, fixed battery exponents have argued that clinicians using a flexible collection of tests are likely to draw incorrect clinical inferences (Russel & Russell, 2003; Russell, Russell & Hill, 2005). These authors outline the apparent vulnerability of flexible battery methodology to:

1. Errors of test selection due to:
    - incomplete testing of all relevant domains;
    - selection of tests specifically to "prove" a favoured hypothesis;
    - administration of tests until evidence is found for a favoured hypothesis;
    - selection of inappropriately easy or difficult tests.

2. Errors of test combination due to:
    - misunderstanding of normal variation between battery scores;
    - lack of co-norming;
    - comparison of tests with non-equivalent normative groups;
    - failure to use appropriate demographic corrections with norms.

3. Failure to empirically evaluate the reliability and validity of the test battery as a whole.

As discussed in chapter one, such exponents are apt to contend that the use of a fixed battery constitutes the most complete application of psychometric principles to the evaluation of cognitive functioning at the battery level (Reitan & Wolfson, 2004; Reitan & Wolfson, 1996; Reitan & Wolfson, 1993; Russell & Russell, 2003; Russell, Russell & Hill, 2005).

The preceding section argues the role of RAPT methodology in providing answers to such challenges for clinicians who employ semi-flexible collections of cognitive tests. RAPT methodology is arguably capable of providing the level of psychometric strength reportedly intrinsic to fixed batteries, to a flexible collection of cognitive tests. In fact, the methodology arguably goes beyond adding value to flexible batteries and could be employed to enhance current fixed battery methodology. Lengthy discussion of the debate between adherents of fixed versus flexible batteries has been avoided thus far due to the high base rate of semi-flexible battery use in clinical practice and the research goal of generating a clinically useful approach to analysing psychological test data. At this point, however, a consideration of the fixed battery in light of the preceding explication of RAPT methodology seems appropriate.

Theoretically, the unchanging nature of the fixed battery facilitates psychometric strength in both normative and ipsative comparisons. For example, the battery may be normed as a whole, eliminating error due to inequality of norms, and relationships between individual measures may be thoroughly evaluated and applied to interpretation. Additionally, an unchanging battery lends itself to empirical validation. It may be validated as a whole according to its ability to achieve the purpose for which it was designed and according to adherence to the cognitive theory upon which it is based. A fixed battery may also be structured according to a specific theory of cognitive functioning (Reitan & Wolfson, 2004; Russell, Russell &

Hill, 2005). Perhaps most importantly, however, a fixed battery readily lends itself to analytical and inferential methodologies, such as composite methodology, which allow an understanding of measurement error and domain validation to play a fundamental role in clinical decision making.

In fact, an unchanging collection of tests is best placed to capitalise on the strengths of psychometrically based inferences and, were these strengths to be optimised in the analytical structures of fixed batteries, fixed battery proponents may be forgiven for wondering why flexible battery users "would not want to take advantage of the important advances in neuropsychological interpretation contributed by multivariate research" (Goldstein, 1997, p. 81).

If, as proponents claim, fixed batteries facilitate inferences based on empirically validated cognitive structure and known reliability then such methodology does indeed constitute the "state of the art". These strengths are highly desirable to the clinician and also constitute fundamental guidelines of the currently proposed RAPT methodology. Closer examination of the construction and analytical framework of common fixed batteries, however, indicates a surprising failure to capitalise on their potential psychometric strengths. This may be examined with specific reference to the fixed battery which clinicians most commonly report using, the HRNB (Butler, Retzlaff & Vanderploeg, 1991; Guilmette, Faust, Hart & Arkes, 1990; Seretny, Dean, Gray & Hartlage, 1986).

The current HRNB originated with the work of Ward Halstead (1947) who compiled a battery of seven tests to identify individuals with frontal lobe lesions through use of a summary impairment index derived from test scores. Halstead's investigation aimed to conceptualise and quantify "behavioural effects of brain lesions in man" (Halstead, 1947, p. vi) and while Halstead was well aware of "modern" techniques such as the derivation of "intelligence quotients" and use of

"factor analysis" he questioned the clinical utility of such methods in quantifying the behavioural correlates of brain damage (p.4). The original Halstead-Reitan Battery (HRB; Halstead, 1947) was modified by the addition of new tests with the specific aim of making neurological diagnoses (Reitan, 1955; Russell, 2000, 1998, 1997, 1995). Three derivative batteries, the Neuropsychological Deficits Scale (Reitan, 1987), which focuses on the identification and lateralization of brain-damage, the Comprehensive Norms for an Extended Halstead-Reitan Battery (CHEHRB; Heaton, Grant & Matthews, 1991) which adds tests to the basic HRNB, and the Halstead Russell Neuropsychological Evaluation System (Russell & Starkey, 1993), which focuses on computerised scoring and also adds subsequent tests to the original HRB, have followed. As an aside, it is interesting to note that this "fixed" battery has several ardently contested versions and numerous modifications reported by practicing clinicians (Lees-Haley, Smith, Williams & Dunn, 1995) raising the question of which of the myriad variants of the HRB is the "fixed" one.

As fixed batteries, the HRNB and derivatives, still lend themselves to empirical validation of structure. Specifically, the constructs underlying measurement for each battery may be evaluated using empirical methodology. The validity of structure-based inferences may be relied upon whenever a battery replicating that administered in validation studies is given. However, while the HRNB is structured according to domains of brain functioning: (1) input; (2) attention, concentration and memory; (3) verbal abilities; (4) spatial, sequential and manipulatory abilities; (5) abstraction, reasoning, logical analysis and concept formation; and (6) output; empirical validation of these constructs has been explicitly avoided in favour of studies of diagnostic validity (Reitan & Wolfson, 1985). The cognitive structure upon which the battery was formulated is not used in analysis or interpretation of tests scores, and does not appear to have been empirically validated.

Both individual tests of the HRNB and overall indices of impairment, such as the Halstead Impairment Index (HII; Reitan & Wolfson, 1985) and the General Neuropsychological Deficit Scale (GNDS; Reitan & Wolfson, 1993, 1988), have demonstrated capacities to discriminate between brain-damaged and non-brain damaged people and to lateralise lesions (Reitan & Wolfson, 1993; Rojas & Bennett, 1995). Furthermore, Reitan and Wolfson (2004) stress the size of HRNB validation samples, arguing that the battery has been validated on "thousands" of patients resulting in definitive validity for diagnoses of both the locality and chronicity of lesions. However, these authors fail to indicate which of the several HRN "batteries" have been subject to this intensive scrutiny. Further, diagnosis is far from the primary aim of cognitive assessment and any amount of validation for the purposes of diagnostic inferences indicates little about the validity of the measure for other, perhaps more pertinent, clinical tasks. Despite the potential of the HRNB to comment on a series of potentially meaningful cognitive domains or modalities, little to no validation of these constructs has occurred. Given that many clinicians wish to evaluate functioning on cognitive domains, and indeed consider test scores as indicative of some underlying factor, construct validation of the HRNB would arguably provide more useful information to practicing clinicians despite the dearth of relevant validity studies.

As indicated in previous discussions, the internal consistency associated with measurement of relevant domains of functioning provides valuable information about the error which is inevitably incorporated into any test scores. Neither the manual of the HRNB (Reitan & Wolfson, 1985) nor that of the CHEHRB (Heaton, Grant & Matthews, 1991) provides any evaluation of internal consistency. Additionally, few studies evaluate the reliability of the HRNB as a whole (Franzen, 2000), none, to my knowledge, investigate internal consistency, and the random error associated with

test scores is universally disregarded in analysis and interpretation of HRNB tests scores (Heaton, Grant & Matthews, 1991; Reitan & Wolfson, 1985). Franzen (2000) reports that while the few psychometric studies conducted indicate only moderate test-retest coefficients (see Dikmen, Heaton, Grant & Temkin, 1999), HRNB proponents have argued that "clinical reliability", as defined by capacity to correctly classify "normal" individuals from those with "impaired" performance, is high for HRNB tests and indices. The capacity to provide consistent diagnoses, however, does not substitute for investigation of the psychometric reliability of test scores and misuse of the term "reliability" in no way constitutes evidence for robust internal consistency. Validity is not sufficient for reliability and the clinician has little choice but to interpret HRNB test score with no direct reference to random error.

Reitan and Wolfson's (1985) recommended analytical methodology makes use of cut-scores which indicate the presence or absence of impairment. Similarly, the analytical structure proposed by Heaton, Grant and Matthews (1991) relies on the conversions of raw scores into T scores which are compared against a cut-off point indicative of impaired performance. Neither of these methodologies is invulnerable to the impact of random errors, which are highly likely to be implicit in any observed score. As discussed in chapter two, reliability directly moderates the cut-scores appropriate for accurate discrimination. That is, the comparison of an individual's observed score to an arbitrary cut-off point is fundamentally confounded by the degree to which the observed score includes error and when reliability is not considered misclassification is likely (Charter & Feldt, 2001). Even given the justifiable confidence with which clinicians may use the HRNB to diagnose the presence of brain damage, ignoring reliability is psychometrically unjustified.

Reviews of clinical practice (such as Lees-Haley, Smith, Williams & Dunn, 1995) strongly suggest that clinicians are more influenced by clinical utility and

practical concerns with regard to their battery construction. As Reitan and Wolfson (1985) accurately state, however, any empirically established validity or reliability and the advantages of known test relationships do not apply when a "fixed" battery is modified by clinicians. In other words a fixed battery only retains its psychometric properties when clinical use duplicates the tests "exactly as they where when the validation studies were done" (Reitan & Wolfson, 1985, p.40). This is not the case in the majority of clinical practice, even when clinicians claim use of a "fixed" battery. What is not recognised by these authors is that modifications to the fixed battery may improve reliability, and warnings of dire consequences need to be empirically examined and not accepted as a matter of faith.

Despite its cited strengths of validity, reliability and clinical applicability, the HRNB addresses few of the theoretical concerns identified by psychometricians as most essential to the accuracy, stability and meaningfulness of test scores. The battery is empirically supported as a diagnostic tool and based on these high levels of validity is likely to demonstrate a corresponding level of stability. However, the valuable information provided by empirical investigation of this psychometric characteristic is lacking. Clinicians seeking to develop an accurate and stable description of cognitive domains using the battery will be impeded by a lack of psychometric information and analytical structure and will, of necessity, draw conclusions fraught with an unknown degree of error. Use of any fixed battery presents challenges to clinicians in terms of clinical utility. When such challenges are not counterbalanced by an improved capacity for psychometrically driven analyses the methodology provides questionable benefits and perhaps contributes little above the capacity of individual tests towards accurate quantification of human cognitive behaviours.

5.3 RAPT Algorithms

The previous section has outlined a rationale for use of RAPT methodology including discussion of the role of RAPT in improving assessment practices using both fixed and flexible collections of cognitive tests. While rational evaluation is useful, empirical proofs of the claimed advantage of the methodology are required. This evaluation will be undertaken in Chapter 6. The following section will present the algebraic formulae which form the "toolbox" for the computation and analysis of composite scores using the individual test scores derived from a semi-flexible testing battery.

5.3.1 Normative Information

For composite calculations, certain normative information is required for each test. An advantage of the composite approach is that the required data are standard features of any standardised test. As discussed above, the applicability of this information to the individual client should be thoroughly evaluated, as should the stability of the reliability coefficient.

Table 5.3

Normative Data Required

For each Test:	**For the Battery:**
Mean Standard Deviation Reliability	Intercorrelations between tests

5.3.2 Equivalence of Scales

Rescaling all composite measures to a single scaling system simplifies computations and avoids errors which occur due to scaling disparities, as discussed in section 4.3.3. Establishing equivalence of test scaling units is the first methodological step in using clinical composites. Linear transformation can be used to produce comparable scales for tests within a battery allowing comparison or combination without changing fundamental relationships between the scores (Anastasi & Urbina, 1997; Murphy & Davidshofer, 2005; Nunnally, 1978).

Composite calculations can be conducted without scaling. However, expressing subtest scores in standard (or standardised) scores makes subsequent composite calculations more "reasonable" (Ghiselli, 1964). As discussed previously, while any scaling system could be employed, the Wechsler scaled score distribution with a population mean of ten and standard deviation of three is chosen in the current method for its likely familiarity to clinicians.

Standardisation of individual test scores requires the normative mean and standard deviation for each test (i.e., based on relevant demographic variables) and use of formula 5.1 or 5.1a below producing a scaled score for an observed score as follows:

$$SS = \left([X - M]/SD\right) \times \sigma_{SS} + \mu_{SS} \qquad \text{Formula 5.1}$$

$$SS = \left([X - M]/SD\right) \times 3 + 10 \qquad \text{Formula 5.1a}$$

Where:

X = the observed score

M = the normative mean of the observed score

SD = the normative standard deviation of the observed score

σ_{SS} *= population standard deviation of the standardisation distribution*

μ_{SS} *= population mean of the standardisation distribution*

As indicated in formula 5.1a, the mean (10) and standard deviation (3) of the scaled score distribution are substituted into the formula when test scores are transformed into scaled scores.

5.2.3 Observed Score

As recommended by Ghiselli (1964), simple additive composites are calculated using the formula for the composite observed score to combine an equally weighted scaled score for each subtest, as follows:

$$X_c = SS_1 + SS_2 + ... + SS_k \qquad \text{Formula 5.2}$$

An empirical relationship, such as those revealed in factor analytic research, should exist between the subtest scaled scores summed to produce the composite observed score. It should be noted that this is a simplification of the general formula in which differential weightings can be applied to each subtest. Simple summation merely applies a weighting of 1.0 to each subtest. This clearly assumes that each measure should contribute equally to the total score, an outcome which is rarely if ever true in practice. This approach has been adopted as the information necessary to apply differential weightings is not generally available to practitioners who utilise psychological tests. This is primarily because weightings, like factor loadings, can vary substantially with different samples and clinical settings. It is worth noting, however, that perhaps the ultimate expression of RAPT methodology would be exemplified by clinicians who construct their test battery using the reliabilities and intercorrelations from their own clinical setting. In such a circumstance the use of weightings would be, not just possible, but recommended.

5.2.4 Composite Mean

The mean of the composite is equal to the sum of the means of the individual components (Ghiselli, 1964).

$$M_c = \mu_{SS\,1} + \mu_{SS\,2} + ... + \mu_{SS\,k} \qquad \text{Formula 5.3}$$

$$M_c = 10(k) \qquad \text{Formula 5.3a}$$

Where:

$\mu_{SS\,k}$ *= the population mean for the standardisation distribution of each Subtest*

k = total number of subtests

The composite mean is therefore a combined expression of the central tendency of the normative group associated with each test. Again, the general formula is simplified to a formula (formula 5.3) in which a weighting of 1.0 is assumed for each subtest mean. As discussed above in relation to the summed observed score, this assumes the equal contribution of measures which is likely to be both theoretically incorrect and practically unavoidable given the typical information available to clinicians. Additionally, as indicated in formula 5.3a, the calculation may be further simplified with the assumption that composites are comprised of subtests represented as scaled scores with a mean of ten.

5.2.5 Composite Standard Deviation

The standard deviation of the composite is equivalent to the sum of the squared deviations of the sample for each subtest, divided by the total number of squared deviations (Ghiselli, 1964).

$$SD_c = \sqrt{\sigma_{SS\,1}^{\,2} + \ldots + \sigma_{SS\,k}^{\,2} + 2\sigma_{SS\,1}\sigma_{SS\,2}r_{12} + \ldots + 2\sigma_{SS\,k}\sigma_{SS\,k-1}r_{k(k-1)}}$$

Formula 5.4

Where:

$\sigma_{SS\,k}^{2}$ = *the population variance of the standardisation distribution of each subtest*

$\sigma_{SS\,k}$ = *the population standard deviation of the standardisation distribution of each subtest.*

$r_{k(k-1)}$ = *the intercorrelation between all subtests included in the composite*

This standard deviation is, therefore, an expression of the average deviation from the mean within the normative sample for each subject and the intercorrelations between subtests within the composite.

In the situation where subtest scores are transformed as recommended to scaled scores with a mean of 10 and a standard deviation of 3, the computational formula for composite standard deviation simplifies and may be computed using formula 5.5 below.

$$SD_c = 3\sqrt{3 + 2\left(\sum r_{k(k-1)}\right)} \qquad \text{Formula 5.5}$$

5.2.6 Composite Reliability

Tellegen and Briggs (1967) commented on the need to specifically evaluate the reliability of any proposed subtest combination and provided the formula to do so with the aim of facilitating clinical use of new subtest combinations such as those suggested by factor analytic research. Reliability of the composite is calculated using the formula below.

$$r_{cc} = \frac{\sum r_{kk'} + 2\sum r_{k(k-1)}}{n + 2\sum r_{k(k-1)}} \qquad \text{Formula 5.6}$$

Where:

$r_{kk'}$ = *the reliability coefficient for each subtest*

As indicated in the formula the reliability coefficient relies on the summed reliabilities of individual tests, the number of tests and the summed test intercorrelations and thus is an expression both of the individual internal consistency of the tests and the degree to which each varies within the domain.

5.2.7 Intercorrelations between Composites

The intercorrelation between composites can be calculated using the following formula (Ghiselli, 1964):

$$r_{c_i c_j} = \frac{\sigma_{SSi_1}\sigma_{SSj_1} r_{i_1 j_1} + \ldots + \sigma_{SSi_l}\sigma_{SSj_m} r_{i_1 j_m} + \ldots + \sigma_{SSi_k}\sigma_{SSj_1} r_{i_k j_1} + \ldots + \sigma_{SSi_k}\sigma_{SSj_m} r_{i_k j_m}}{SD_{c_i} SD_{c_j}}$$

Formula 5.7

Where:

$\sigma_{SS\,i}$ = *population standard deviation of the standardisation distributions of subtests from composite "i"*

$\sigma_{SS\,j}$ = *population standard deviation of the standardisation distributions of subtests from composite "j"*

r_{ij} = *intercorrelations between subtests in composite "i" and subtests in composite "j"*

SD_{ci} =*standard deviation of composite "i"*

SD_{cj} =*standard deviation of composite "j"*

Calculation of composite intercorrelations is required for the evaluation of abnormal discrepancies between composites and requires a matrix of intercorrelations between all subtests in both composites, all subtest standard deviations and the standard deviation of each composite.

5.2.8 Deviation Quotient

Deviation Quotient is calculated using a simple linear transformation, based on a distribution with a mean of 100 and a standard deviation of 15. As discussed in the previous section while this calculation would simplify mathematically and conceptually if based on z-scores, the familiarity of the standard score distribution warrants usage. Thus, as with individual subtest scaled scores, the standard score distribution is chosen for its likely familiarity to the clinician (Cicchetti, 1994) and is calculated using the following formulae.

$$DQ_c = \left((X_c - M_c)/SD_c\right) \times \sigma_{ST} + \mu_{ST} \qquad \text{Formula 5.8}$$

$$DQ_C = \left((X_C - M_C)/SD_C\right) \times 15 + 100 \qquad \text{Formula 5.8a}$$

Where:

σ_{ST} = *population standard deviation of the standard score*

μ_{ST} = *population mean of the standard score*

As indicated in formula 5.8a, the standard score mean of 100 and standard deviation of 15 are incorporated into the calculation of a linear transformation of the summed composite observed score. Once the deviation quotient is calculated this standardised score may be used to determine the percentile ranking of the individual under the normal curve.

5.2.9 Confidence Intervals

Clinicians aim to measure a client's true ability on a task. Despite this, psychological reports tend to provide the only observed scores gained during the course of testing. The reliability or stability coefficients of psychological tests are the clinician's guide to the magnitude of random error variance in a client's

performance. The higher the reliability of the test, the less the error variance, and the more likely it is that the observed score is representative of the client's true ability, without the impact of random error variance. Psychologists have customarily dealt with the uncertainty resulting from less than perfect test reliability by examining test scores within a confidence band or interval, which is recommended in the current methodology.

The standard error of estimate forms the basis for confidence intervals centred on the predicted true score (as discussed in section 2.3.2). Calculation of the standard error provides a practical indication of the degree to which error is likely to impact on the observed test score as it relies directly on the composite reliability. The standard error of estimate, SE_E, (formula 2.5) is used to establish confidence intervals for test scores based on the reliability of the composite (r_{cc}; formula 5.6). The predicted true score plus or minus the SE_E (formula 2.6) encompasses the interval in which two-thirds of scorers with the true score will lie.

$$SE_{E_c} = \sigma_{ST}\sqrt{(r_{cc})(1-r_{cc})} \qquad \text{Formula 5.9}$$

$$SE_{E_c} = 15\sqrt{(r_{cc})(1-r_{cc})} \qquad \text{Formula 5.9a}$$

Formula 5.9a is the simplified formula for composites based on a standard score distribution (i.e., $\mu_{ST}=100$; $\sigma_{ST}=15$).

The confidence interval is produced by multiplying the z score corresponding to a particular level of confidence by the standard error of estimate. A 90% confidence level, corresponding to a z-score of 1.64, reflects a good compromise between the need for a high degree of confidence and intervals that still have clinical utility. The confidence interval describes a band of scores calculated using formula 5.10 and corresponding to the proscribed level of confidence.

$$DQ_{PT_c} \pm (z) * \left(SE_{E_c}\right) \qquad \text{Formula 5.10}$$

Where:

Confidence Level	68%	85%	90%	95%	99%
z-score	1.00	1.44	1.64	1.96	2.58

The confidence interval around the observed score DQ indicates the band within which the actual score would be expected to fall a certain percentage of the time (e.g., 90%, 95%). Again, as discussed in chapter two, the confidence interval is most accurately centred around the predicted true score which provides an estimate of true score accounting for regression to the mean (Dudek, 1979; Glutting, McDermott & Stanley, 1987). The confidence band is placed symmetrically around the predicted true score (computed using formula 5.11, or 5.11a when a standard score distribution is assumed) resulting in an asymmetric placement around the observed score (formulae 5.8 or 5.8a). The predicted true score makes use of the reliability estimate associated with the score to moderate for potential regression to the mean, as follows:

$$DQ_{PTc} = r_{cc}\left(DQ - \mu_{ST}\right) + \mu_{ST} \qquad \text{Formula 5.11}$$

$$DQ_{PTc} = r_{cc}\left(DQ - 100\right) + 100 \qquad \text{Formula 5.11a}$$

The third standard error, the standard error of prediction, is used to establish confidence intervals that encompass two-thirds of the scores upon retest, given a true score at test. While SE_P efficiently accommodates for measurement error, based on test reliability, some authors have noted the inability of this index to account for practice effects (Glutting, McDermott & Stanley, 1987). This should be considered when the resulting confidence intervals are interpreted. The formula below (formula 5.12) is again modified (formula 5.12a) with the assumption that standardised scores are used throughout the RAPT methodology and as with the previous standard error formula is based on a variation on the composite reliability.

$$SE_{Pc} = \sigma_{ST}\sqrt{(1 - r_{cc}{}^2)} \qquad \text{Formula 5.12}$$

$$SE_{Pc} = 15\sqrt{(1 - r_{cc}{}^2)} \qquad \text{Formula 5.12a}$$

The re-test confidence interval is produced by multiplying the z score corresponding to a particular level of confidence by the standard error of prediction. This reflects a slight modification on formula 5.10 (formula 5.10a), as follows:

$$DQ_{PT_c} \pm (z) * \left(SE_{P_c}\right) \qquad \text{Formula 5.10a}$$

5.2.11 Significance of Composite Differences

A difference between composite scores is "significant" if it is so large that it is unlikely to have occurred due to chance or measurement error (Silverstein, 1981). Evaluation of this discrepancy is fundamentally reliant on the intercorrelation between the compared measures. If two measures are highly correlated and if two composites are comprised of highly correlated subtests, then even small discrepancies are unlikely to have occurred due to chance (Payne & Jones, 1957). Evaluation of a significant difference also relies upon the reliability of the measures compared. As test reliability decreases, the likelihood that observed discrepancies may have occurred by chance increases (Payne & Jones, 1957).

The standard error of measurement (formula 5.13), reflects the error variance for the standardisation sample and should only be used for confidence intervals if the client is a legitimate member of this sample. The standard error of measurement is, in fact, larger than the more appropriate standard error of estimate and therefore generates confidence bands that are larger than is in fact the case. The SE_M is necessary, however, to compute the standard error of the difference between two composites (formula 5.14). Again, the standard deviation of the composite

population is used in the formula along with the composite reliability, and the formula modified for use with standard scores.

$$SE_{Mc} = \sigma_{ST}\sqrt{1 - r_{cc}} \qquad \text{Formula 5.13}$$

$$SE_{Mc} = 15\sqrt{1 - r_{cc}} \qquad \text{Formula 5.13a}$$

Evaluation of the significance of a discrepancy between two composite scores is obtained by dividing the difference between the two scores by the standard error of measurement of the difference. This standard error is the denominator of formula 5.14 below (Silverstein, 1981):

$$Z_S = \frac{D_{ij}}{\sqrt{SE_{Mi}{}^2 + SE_{Mj}{}^2}} \qquad \text{Formula 5.14}$$

Where:

D_{ij} = Difference between the Deviation Quotients of composite "i" and "j"

SE_{Mi} = the standard error of measurement associated with composite "i"

SE_{Mj} = the standard error of measurement associated with composite "j"

The resulting value may be referred to the standard normal distribution. More conveniently, the difference necessary for significance at a specific level of z may be determined by multiplying the standard error of the measurement of the difference (the result of the denominator of formula 5.14) by the appropriate z value (e.g. 1.96 for $p < .05$ level). This will provide a cut-of level for "significance".

5.2.12 Abnormal Differences Between Composite Scores

Determining the presence of abnormal variation between composite scores potentially presents the more important computational procedure. Clinicians may be equally concerned with meaningful differences (i.e., in addition to consideration of significant differences). Critical abnormality levels are calculated by multiplying the standard deviation of the difference between two scores (the denominator of formula 5.15) by the unit normal deviate corresponding to the desired frequency level. Note

that this value is two-tailed so for a unidirectional test you would use a correspondingly more rigorous value (i.e., 2.5%). Estimation of the abnormality of the discrepancy, "D_{ij}" between two scores in the normal population then involves looking up the frequency associated with the z-score calculated using the following formula (Silverstein, 1981):

$$Z_d = \frac{D_{ij}}{\sqrt{\sigma_{STi}^2 + \sigma_{STj}^2 - 2r_{c_i c_j}\sigma_{STi}\sigma_{STj}}} \qquad \text{Formula 5.15}$$

Where:

σ_{STi} = population standard deviation of the standard score for composite "i"

σ_{STj} = population standard deviation of the standard score for composite "j"

Normative Frequency	10%	5%	1%
z-score	1.28	1.64	2.32

As indicated in the formula, the difference between composites is divided by the standard deviation of difference (i.e., the formula denominator, which is a function of the individual composite standard deviations and intercorrelations) and associated with the appropriate frequency. Alternately, the standard deviation of the difference may be multiplied by the required z-value (e.g., 1.28 for 10%) to indicate the cut-off score above which a discrepancy is abnormal.

5.4 Demonstration of RAPT Methodology

In the following section the RAPT methodology (outlined in section 5.2) and the algorithms (outlined in section 5.3) will be demonstrated using two composite examples. For simplicity, the composites calculated will replicate the Verbal Comprehension Index (VCI) and the Working Memory Index (WMI) of the WAIS-

III. These examples are suitable for several reasons: first their familiarity to the clinician and second the ready availability of the requisite normative and psychometric information for each composite subtest. Calculations will make use of a specific case example with the aim of demonstrating sequential use of the RAPT methodology and algorithms to combine and analyse test scores and draw clear clinical conclusions regarding the functioning of the individual case. Despite the seeming complexity of the previous algorithms, RAPT methodological steps facilitate this process efficiently. The following demonstration highlights the ease with which RAPT methodology may be applied to clinical tasks using the following three stages.

5.4.1 Stage One

During stage one the battery is structured and normative data and raw scores collected. Both the VCI and WMI are based on clearly validated structure as indicated in the WAIS-III technical manual (Wechsler, 1997a). Subtests to be included in the VCI are Vocabulary (VOC), Information (INF) and Similarities (SIM). Those included in the WMI are Arithmetic (AR), Digit Span (DSp) and Letter-Number Sequencing (LNS). The correlation matrix for each subtest is accessed from the WAIS-III and WMS-III technical manual (Wechsler, 1997a) and presented in Table 5.4. The subtest scores are WAIS-III scaled scores with equal means of 10 and standard deviations of 3. No transformation to a similar scaling system is required. Table 5.5 provides the example scaled scores and the reliability coefficients associated with the subtests.

Table 5.4

Correlation Matrix for VCI and WMI Subtests

	AR	LNS	DSp	VOC	INF
LNS	.55				
DSp	.52	.57			
VOC	.60	.50	.45		
INF	.63	.47	.40	.77	
SIM	.57	.46	.40	.76	.70

Table 5.5

Scaled Scores and Reliabilities for VCI and WMI Subtests

	WMI			VCI		
	AR	**LNS**	**DSp**	**VOC**	**INF**	**SIM**
Scaled Scores	9	8	7	3	4	5
Reliability	.88	.90	.82	.93	.91	.86

5.4.2 Stage Two

Stage two involved analysis of the collected raw and normative data. The composite observed scores, composite means, composite standard deviation and composite reliability are calculated using formulae 5.2, 5.3a, 5.4, 5.5 and 5.6 respectively, as indicated in figure 5.5. This figure provides a brief description of these procedures as well as the formulae and calculations for each step.

5.4.3 Stage Three (Part A)

The first aspect of stage three involves normative comparisons. The composite deviation quotients (DQ; M=100; SD=15) are calculated using formula 5.8a along with associated percentiles. Both the standard error of estimate (formula 5.9a) and standard error of measurement (formula 5.13a) are calculated as well as the predicted true deviation quotient for use in confidence intervals (formula 5.11a). Finally, ninety percent confidence intervals are calculated using the standard error of estimate, DQ_{PT} and formula 5.10. Calculations are demonstrated in figure 5.6 which again outlines both the rationale and the calculations occurring at this stage.

5.4.4 Stage Three (Part B)

Finally, the last aspect of stage three involves ipsative comparisons. A comparison between the VCI and WMI composites is conducted to evaluate both the significance (formula 5.14) and abnormality (formula 5.15) of the discrepancy between VCI and WMI, as indicated in Figure 5.7. As indicated in the figures below, the deviation quotient for working memory is 88 (i.e., 83 – 95) indicating functioning in the Low Average range, and for verbal comprehension is 67 (i.e., 64-74) indicating functioning in the Extremely Low range, according to WAIS-III classifications. These scores are both significantly and abnormally discrepant. Specifically, the 21 point difference between the WMI and VCI composites is significant at the .05 level and this difference would be expected to occur in the most extreme 10% and 5% of the population.

Figure 5.2 Stage Two calculations: composite observed score, mean, standard deviation and reliability and intercorrelation between composites

Formulae	**WMI**	**VCI**
2.1 Composite Observed Score		
$X_c = SS_1 + SS_2 + ... + SS_k$	**SS_{AR}** = 9 **SS_{LNS}** = 8 **SS_{DSp}** = 7	**SS_{VOC}** = 3 **SS_{INF}** = 4 **SS_{SIM}** = 5
Validated structure is consolidated by summing the observed scores for tests of each domain.	$X_{WMI} = 9 + 8 + 7$ $X_{WMI} = 24$	$X_{VCI} = 3 + 4 + 5$ $X_{WMI} = 12$
2.2 Composite Mean		
$M_c = 10(k)$	**M_{AR}** = 10 **M_{LNS}** = 10 **M_{DSp}** = 10	**M_{VOC}** = 10 **M_{INF}** = 10 **M_{SIM}** = 10
The normative central tendency of each normative sample is expressed in combination for the composite by an additive composite	$M_{WMI} = 10(3)$ $M_{WMI} = 30$	$M_{VCI} = 10(3)$ $M_{VCI} = 30$

Figure 5.2 (continued)

Formulae	**WMI**	**VCI**
2.3 Composite Standard Deviation		
$SD_c = \sqrt{\sigma_{SS1}^2 + \ldots + \sigma_{SSk}^2 + 2\sigma_{SS1}\sigma_{SS2}r_{12} + \ldots + 2\sigma_{SSk}\sigma_{SSk-1}r_{k(k-1)}}$	**SD_{AR}** = 3 **V_{AR}** = 9 **SD_{LNS}** = 3 **V_{LNS}** = 9 **SD_{DSp}** = 3 **V_{DSp}** = 9	**SD_{VOC}** = 3 **V_{VOC}** = 9 **SD_{INF}** = 3 **V_{INF}** = 9 **SD_{SIM}** = 3 **V_{SIM}** = 9
$SD_c = 3\sqrt{3 + 2\left(\sum r_{k(k-1)}\right)}$	**$r_{AR\ LNS}$** = .55 **$r_{AR\ DSp}$** = .52 **$r_{LNS\ DSp}$** = .57	**$r_{VOC\ INF}$** = .77 **$r_{VOC\ SIM}$** = .76 **$r_{INF\ SIM}$** = .70
Normative deviations characteristic of each test, coupled with the correlational relationships shared by tests provide an expression of deviation for the composite.	$SD_{WMI} = \sqrt{3^2 + 3^2 + 3^2 + 2(3)(3)0.55 + 2(3)(3)0.52 + 2(3)(3)0.57}$	$SD_{VCI} = \sqrt{3^2 + 3^2 + 3^2 + 2(3)(3)0.77 + 2(3)(3)0.76 + 2(3)(3)0.70}$
	$SD_{WMI} = \sqrt{9 + 9 + 9 + (18)0.55 + (18)0.52 + (18)0.57}$	$SD_{VCI} = \sqrt{9 + 9 + 9 + (18)0.77 + (18)0.76 + (18)0.70}$
	$SD_{WMI} = \sqrt{27 + 9.9 + 9.36 + 10.26}$	$SD_{VCI} = \sqrt{27 + 13.86 + 13.68 + 12.6}$
	$SD_{WMI} = 7.5$	$SD_{VCI} = 8.2$
	Alternately the simplified formula may be used:	
	$SD_{WMI} = 3\sqrt{3 + 2\left(\sum 0.55 + 0.52 + 0.57\right)}$	$SD_{VCI} = 3\sqrt{3 + 2\left(\sum 0.77 + 0.76 + 0.70\right)}$
	$SD_{WMI} = 3\sqrt{3 + 2(1.64)}$	$SD_{VCI} = 3\sqrt{3 + 2(1.49)}$
	$SD_{WMI} = 7.5$	$SD_{VCI} = 8.2$

Figure 5.22 (continued)

Formulae	**WMI**		**VCI**	
2.4 Composite Reliability				
$r_{cc} = \frac{\sum r_{kk'} + 2\sum r_{k(k-1)}}{n + 2\sum r_{k(k-1)}}$	$\mathbf{r_{AR\ LNS}} = .55$ $\mathbf{r_{AR\ DSp}} = .52$ $\mathbf{r_{LNS\ DSp}} = .57$	$\mathbf{r_{AR}} = .88$ $\mathbf{r_{DSp}} = .90$ $\mathbf{r_{LNS}} = .82$	$\mathbf{r_{VOC\ INF}} = .77$ $\mathbf{r_{VOC\ SIM}} = .76$ $\mathbf{r_{INF\ SIM}} = .70$	$\mathbf{r_{VOC}} = .93$ $\mathbf{r_{SIM}} = .86$ $\mathbf{r_{INF}} = .91$
Test interrelationships and reliabilities are used to calculate the internal consistency coefficient for composites.	$r_{WMI} = \frac{(0.88+0.90+0.82)+2(0.55+0.52+0.57)}{3+2(0.55+0.52+0.57)}$ $r_{WMI} = \frac{2.6+3.28}{3+3.28}$ $r_{WMI} = \frac{5.88}{6.28}$ $r_{WMI} = 0.94$		$r_{VCI} = \frac{(0.93+0.86+0.91)+2(0.77+0.76+0.70)}{3+2(0.77+0.76+0.70)}$ $r_{VCI} = \frac{2.7+4.46}{3+4.46}$ $r_{VCI} = \frac{7.16}{7.46}$ $r_{VCI} = 0.96$	

Figure 5.2 (continued)

Formulae	WMI	VCI

2.4 Intercorrelation between Composites

$$r_{c_i c_j} = \frac{\sigma_{SSi_1}\sigma_{SSj_1} r_{i_1 j_1} + \ldots + \sigma_{SSi_j}\sigma_{SSj_m} r_{i_1 j_m} + \ldots + \sigma_{SSi_k}\sigma_{SSj_1} r_{i_k j_1} + \ldots + \sigma_{SSi_k}\sigma_{SSj_m} r_{i_k j_m}}{SD_{c_i} SD_{c_j}}$$

Test standard deviations and the standard deviation for the composite combined to indicate the degree to which two composite correlate.

SD_{AR} = 3 **SD_{WMI}** = 7.5 **SD_{VOC}** = 3 **SD_{VCI}** = 8.2

SD_{LNS} = 3 **SD_{INF}** = 3

SD_{DSp} = 3 **SD_{SIM}** = 3

	AR	LNS	DSp
VOC	.60	.50	.45
INF	.63	.47	.40
SIM	.57	.46	.40

$$r_{WMIVCI} = \frac{(9)(0.60)+(9)(0.50)+(9)(0.45)+(9)(0.63)+(9)(0.47)+(9)(0.40)+(9)(0.57)+(9)(0.46)+(9)(0.40)}{(7.5)(8.2)}$$

$$r_{WMIVCI} = \frac{5.4+4.5+4.05+5.67+4.23+3.6+5.13+4.14+3.6}{61.5}$$

$$r_{WMIVCI} = \frac{40.32}{61.5}$$

$$r_{WMIVCI} = 0.66$$

Figure 5.3 Stage Three (A) calculations: composite deviation quotient (DQ), standard error of estimate (SE_E), standard error of measurement (SE_M), predicted true deviation quotient (DQ_{PT}), observed score and ninety percent confidence intervals

Formulae	**WMI**	**VCI**
3a.1 Composite Deviation Quotient		
$DQ_C = \left((X_C - M_C) / SD_C\right) \times 15 + 100$	$\mathbf{X_{WMI}} = 24$ $\mathbf{M_{WMI}} = 30$ $\mathbf{SD_{WMI}} = 7.5$	$\mathbf{X_{VCI}} = 12$ $\mathbf{M_{VCI}} = 30$ $\mathbf{SD_{VCI}} = 8.2$
Linear transformation to a standard score distribution with mean = 100 and standard deviation = 15 is applied to the additive composite observed scores.	$DQ_{WMI} = \left(24 - 30 / 7.5\right) \times 15 + 100$ $DQ_{WMI} = -0.8 \times 15 + 100$ $DQ_{WMI} = 88$	$DQ_{VCI} = \left(12 - 30 / 8.2\right) \times 15 + 100$ $DQ_{VCI} = -2.2 \times 15 + 100$ $DQ_{VCI} = 67$
3a.2 Standard Error of Estimate		
$SE_{E_c} = 15\sqrt{(r_{cc})(1 - r_{cc})}$	$\mathbf{r_{WMI}} = .94$	$\mathbf{r_{VCI}} = .96$
An index of error based on composite reliability is calculated for use in ipsative comparisons (i.e., the significance of deviations between composites).	$SE_{EWMI} = 15\sqrt{[(0.94)(1 - 0.94)]}$ $SE_{EWMI} = 15\sqrt{0.056}$ $SE_{EWMI} = 3.56$	$SE_{EVCI} = 15\sqrt{[(0.96)(1 - 0.96)]}$ $SE_{EVCI} = 15\sqrt{0.038}$ $SE_{EVCI} = 2.94$

Figure 5.3 (continued)

Formulae	WMI	VCI
3a.3 Standard Error of Measurement		
$SE_{Mc} = 15\sqrt{1 - r_{cc}}$ *An index of error based on composite reliability is calculated for use in calculating confidence intervals.*	**r_{WMI}** = .94 $SE_{MWMI} = 15\sqrt{1 - 0.94}$ $SE_{MWMI} = 15(0.25)$ $SE_{MWMI} = 3.67$	**r_{VCI}** = .96 $SE_{MVCI} = 15\sqrt{1 - 0.96}$ $SE_{MVCI} = 15(0.2)$ $SE_{MVCI} = 3$
3a.4 Predicted True Deviation Quotient		
$DQ_{PTc} = r_{cc}(DQ - 100) + 100$ *Confidence intervals will be centred on this quotient, which is modified according to composite reliability to account for regression to the mean.*	**r_{WMI}** = .94 **DQ_{WMI}** = 88 $DQ_{PTWMI} = 0.94\ (88 - 100) + 100$ $DQ_{PTWMI} = 88.72$ $DQ_{PTWMI} = 89$	**r_{VCI}** = .96 **DQ_{VCI}** = 67 $DQ_{PTVCI} = 0.96(67 - 100) + 100$ $DQ_{PTVCI} = 68.32$ $DQ_{PTVCI} = 69$

Figure 5.3 (continued)

Formulae	**WMI**		**VCI**	
3a.5 Confidence Interval				
$DQ_{PT_c} \pm (z) * (SE_{E_c})$	**$DQ_{PT\ WMI}$** = 89	**z-score**	**$DQ_{PT\ VCI}$** = 69	**z-score**
	$SE_{M\ WMI}$ = 3.67	90% CI = 1.64	**$SE_{M\ VCI}$** = 3	90% CI = 1.64
Confidence intervals provide a band within which the individual's true score is likely to fall.	$89 \pm 1.64(3.67)$		$69 \pm 1.64(3)$	
	89 ± 6.02		69 ± 4.92	
	90% Confidence Interval = 83-95		90% Confidence Interval = 64-74	
	Conclusions: The individual's true score is 90% likely to fall between a deviation quotient of 83 and 95, given the known reliability of measurement.		*Conclusions:* The individual's true score is 90% likely to fall between a deviation quotient of 64 and 74, given the known reliability of measurement.	

Figure 5.4 Stage Three (B) calculations: significance of WMI-VCI discrepancy and abnormality of WMI-VCI discrepancy

3b.1 Significance of WMI-VCI Discrepancy

$$Z_S = \frac{D_{ij}}{\sqrt{SE_{Mi}{}^2 + SE_{Mj}{}^2}}$$

Note: denominator of this formula is the standard error of the difference between scores (SEd).

Formalised means of determining the significance of deviations between composites.

WMI – VCI Discrepancy (D) = 88-67 = 21
$SE_{M\ WMI}$ = 3.67
$SE_{M\ VCI}$ = 3

$$SEd = \sqrt{3.67^2 + 3^2}$$

$$SEd = 4.74$$

The difference needed for significance at .05 is calculated by multiplying the SEd by the z-score associated with this cut-off (1.96):
Difference =1.96(4.74)
= 9.30

On the other hand, the probability associated with the specific discrepancy (21) may be determined by looking up the z-score calculated using formula 5.14 against a table of the normal distribution, as follows:

$$Z = \frac{21}{\sqrt{3.67^2 + 3^2}}$$

$$Z = \frac{21}{4.74}$$

$Z = 4.43$, p-value =<.0001

Conclusions: The difference between the WMI and VCI composites is significant at the .05 level and the specific probability associated with the discrepancies is 0. The individual's performance in these composites is significantly discrepant.

Figure 5.4 (continued)

3b.2 Abnormality of WMI-VCI Discrepancy

$$Z_d = \frac{D_{ij}}{\sqrt{\sigma^2_{STi} + \sigma^2_{STj} - 2r_{c_i c_j}\sigma_{STi}\sigma_{STj}}}$$

Note: denominator of this formula is the standard deviation of the difference between scores (SDd).

Formalised means of determining the frequency of discrepancies between composites likely in the normal population is based on the standard score distribution of the DQ and the intercorrelation between composites.

WMI – VCI Discrepancy (D) = 21

σ_{WMI} = 15

σ_{VCI} = 15

$r_{WMI\ VCI}$ = .66

$$SDd = \sqrt{15^2 + 15^2 - 2(0.66)(15)15}$$
$$SDd = \sqrt{225 + 225 - 297}$$
$$SDd = \sqrt{153}$$
$$SDd = 12.4$$

The difference which would be found in less than or equal to 10% of the population is calculated by multiplying the SDd by the z-score associated with this cut-off (1.28):
Difference = 1.28(12.4)
= 15.83

The difference which could be expected in the most extreme 5% of the population is calculated by multiplying the SDd by the z-score associated with this cut-off (1.64):
Difference =1.64(12.37)
= 20.28

Conclusions: The difference between the WMI and VCI composites is greater than that which could be expected to occur in the most extreme 10% and 5% of the population.

5.5 Conclusions

RAPT provides a systematic methodology of integrating psychometric theory into the use of a semi-flexible collection of cognitive measures. While there is nothing unique or controversial about the psychometric principles or formulae utilised, it is their combination into a systematic approach to psychological test analysis that constitutes the fundamental contribution of the method. The methodology applies psychometric theory to the tasks of selecting, structuring, combining, analysing and interpreting psychometric tests. In this way, while the actual semi-flexible battery is not validated as a whole the psychometric strength of the battery is well known to the clinician and the reliability associated with domain-based test scores is known and used in clinical decision making. Such methodology constitutes an opportune improvement to the most common method in which cognitive tests are used clinically, that is in a changing collection.

RAPT methodology provides clinicians with concrete answers to various challenges posed by professions with which psychologists work in close conjunction. For example, using RAPT methodology the clinician may readily answer questions regarding the validity of inferences about overall cognitive domains and the confidence with which such conclusions are drawn. The RAPT provides flexible battery users with the psychometric strengths traditionally associated with use of a fixed battery and in fact arguably transcends the capacity of commonly used fixed batteries, such as the HRNB. When cognitive test results are used for practical purposes, such as evaluating the extent of cognitive degeneration, planning rehabilitation following injury, identifying the cause of educational impairment, or describing the cognitive strengths and weaknesses of an individual, the capacity to understand the degree to which error impacts upon test scores is vital to the clinician.

Of course, while various logical arguments have been forwarded for the strengths of RAPT methodology, empirical evaluation of the model is required. Such investigations serve RAPT methodology, as studies of empirical validity and reliability do fixed batteries. The results presented in the next chapter will outline empirical evidence of the validity, reliability and utility of using RAPT in clinical practice.

CHAPTER SIX

EVALUATION OF THE RAPT METHDOLOGY

6.1 Introduction

The previous chapter presented the RAPT approach including the algebraic formulae and an explication of the RAPT model. Additionally, a rationale for the applicability of RAPT to clinical practice was presented. Both aspects of the RAPT methodology have been developed with consideration of reliability theory, validity theory and clinical concerns, such as client characteristics and the suitability of norms. This is intended to "formalise" the application of such theory to the structure, analysis and interpretation of a semi-flexible multi-measure battery.

Empirical support is required for any analytical methodology before clinical use is appropriate. To this end, the primary goal of the current chapter is to explore the capacity of the RAPT methodology to evaluate the influence of error in cognitive battery usage. The capacity of composite methodology to provide a rationale for battery composition and analysis will be explored in terms of likely improvements to the reliability and validity of clinical inferences and the usefulness of the battery as a whole. Specifically, composite scores will be demonstrated to provide a means of: measuring validated structure; improving the reliability of domain-based score combinations; providing an empirical rationale for the number of tests needed; and minimising the impact of artifactual error. The capacity of RAPT to assist clinicians to integrate the competing psychometric and clinical demands of battery usage will also be discussed. The current chapter aims to provide empirical evidence of the several strengths of composite methodology as employed by RAPT.

6.2 The Benefits of Composite Scores

Composite scores are a widely used methodology in psychometric assessment, perhaps because of the several advantages gained by combining individual test scores. RAPT methodology aims to capitalise on various strengths of composite scores, as indicated below:

1. Composite scores provide a "formalised" means of measuring validated cognitive domains and analysing and interpreting tests of the same construct in relation to each other. Further, use of composite scores facilitates comparisons across domains using actuarial methods.

2. Composites may improve the stability of construct-based inferences beyond that of any individual test score and can also provide a rationale for test selection.

3. Composite scores diminish the impact of artifactual errors and discriminate more accurately between "impaired" and "non-impaired" performance.

4. Finally, composite scores may be applied to a flexible collection of tests, to facilitate psychometric evaluation and provide actuarial analytical structure. Clinicians may employ RAPT methodology to draw psychometrically robust using clinically useful semi-flexible batteries.

The benefits of composites can be empirically investigated with specific reference to their role in operationalising the structure of semi-flexible test batteries. A myriad of test instruments are available to clinicians and must be combined based

on some systematic, and preferably empirically based, structure of cognitive functioning. RAPT methodology provides a means by which clinicians may evaluate the impact upon reliability of combining tests that underlie empirically validated constructs of cognitive functioning.

6.2.1 Composites Formalise Validated Structure

Composites provide a means of formally applying knowledge about cognitive domains to test combination and analysis. At the most rigorous level of application, test combinations leading to composites may be directly based on the results of factor analytic research. This will be illustrated with a principal components analysis (PCA) conducted on a battery of tests administered as part of ongoing normative studies through the Department of Psychology at the University of Southern Queensland, Australia.

As discussed in chapter three, PCA is an exploratory factor analytic technique designed to empirically summarise a set of variables into a parsimonious number of constructs, that account for the maximum amount of variance within the larger data set (Thompson, 2004). While PCA is far from the only available exploratory method, it is generally considered to produce components which tend to differ only superficially from those derived by more mathematically complex models, such as principal axes factor analysis (Velicer & Jackson, 1990; Gorsuch, 2003). In the current analyses, PCA was chosen for psychometric soundness and mathematical simplicity (Goh, 2006). The chosen method of rotation (promax) facilitated oblique factors and was selected to allow for realistically correlated components and to facilitate simplicity of structure (Cattell, 1978; Thompson, 2004).

Factors with eigenvalues greater than one were retained, a criterion likely to result in over-retention of components (Zwick & Velicer, 1982). Components were

evaluated in terms of their meaning and the results of PCA used to provide suggestions regarding test combination. Thus, for the purpose of the current investigation, the retention of additional components was considered a more desirable outcome than rejecting potentially clinically meaningful test combinations.

The normative sample used for analyses was drawn from a large archival pool of 1045 participants gathered as part of on-going normative studies at the University of Southern Queensland. Given the nature of the data set some variables included in the analysis had missing data as the sample is the combined result of several normative studies which did not always utilise exactly the same combinations of tests. During analyses, pairwise deletion was chosen to accommodate for these missing data: the actual numbers of participants on which correlation matrices are based are indicated in Table 6.1 below. As indicated in the table all of the variables had sufficient cases to satisfy the recommended heuristic of ten cases per variable and the majority of variables had more than 300 cases upon which to base correlations.

Table 6. 1

Number of Normative Cases Analysed for each Variable in the PCA Sample

Variable	Number Analysed
WAIS3-VO	598
WAIS3-SI	598
COWAT	591
STW	537
ANIMALS	522
SDMT-Oral	508
SDMT-Written	507
BNT	441
TMT-Part A	416
TMT-Part B	414
WAIS3-IN	390
WRAT3-Reading	386
WAIS3-LNS	380
WAIS3-AR	379
WAIS3-DSp	378
STROOP-Colour/Word	367
STROOP-Colour	367
STROOP-Word	366
WAIS3-SS	286
WAIS3-CO	284
WAIS3-DSy	283

The normative sample consisted of 431 male and 614 female adults from the general community who had volunteered to participate in studies designed to establish Australian norms for a number of psychological tests. All participants were residents of rural South-East Queensland or metropolitan Brisbane. The mean age of the sample was 34.97 years (SD =14.38) with ages ranging from 16 to 86. The mean

number of years of education was 12.73 (SD = 2.35) and ranged from 5 to 22 years. The test performances included in the PCA were:

- Information (IN), Vocabulary (VO), Similarities (SI), Comprehension (CO), Digit Symbol-Coding (DSy), Symbol Search (SS), Digit Span (DSp), Arithemetic (AR), and Letter Number Sequencing (LNS) subtests of the WAIS-III (Wechsler, 1997a, 1997b)
- the Spot-the-Word (STW) subtest from the Speed and Capacity of Language Processing (SCOLP) test (Baddeley, Emslie & Nimmo-Smith, 1992)
- Written and Oral trials of the Symbol Digit Modalities Test (SDMT) (Smith, 1982)
- the Boston Naming Test (BNT) (Kaplan, Goodglass & Weintraub, 1983)
- the Reading subtest of the third edition of the Wide Range Achievement Test (WRAT-3) (Wilkinson, 1993)
- Trials A and B of the Trail Making Test (TMT) (Charter, Adkins, Alekoumbides & Seacat, 1987)
- Colour, Word, and Colour/Word trials of the Stroop Neuropsychological Screening Test (STROOP) (Golden, 1978)
- Controlled Oral Word Association test (COWAT) (Spreen & Strauss, 1998)
- Animal Naming subtest (ANIMALS) of the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983)

This data was subjected to a PCA with promax rotation. The KMO measure of sampling adequacy was .81 indicating that the data was highly factorable. Five components were retained using the “eigenvalues > 1” rule with factor loadings as

indicated in Table 6.2. For ease of interpretation only component loadings of greater than .5 are displayed in this table.

Table 6. 2

Principle Components Analysis of an Australian Normative Sample

	Component				
Measures	**1**	**2**	**3**	**4**	**5**
WAIS3-IN	.852				
BNT	.804				
WAIS3-VO	.801				
WAIS3-CO	.728				
WAIS3-SI	.711				
Spot the Word	.655				
WRAT3-Reading	.650				
SDMT-Written		.874			
SDMT-Oral		.807			
WAIS3-SS		.777			
TMT-Part A		-.730			
TMT-Part B		-.682			
WAIS3-DSy		.627			
STROOP-Colour			.898		
STROOP-Word			.787		
STROOP-Colour/Word			.599		
WAIS3-LNS				.816	
WAIS3-DSp				.814	
WAIS3-AR				.578	
COWAT					.767
ANIMALS					.617

Examination of the patterns of loadings suggested that the first component could be characterised as a measures of word knowledge (WK) composed of four

verbal subtests from the WAIS-III, the BNT, STW from the SCOLP and Reading, from the WRAT-3. The second component could be characterised as measuring processing speed (PS) composed of Written and Oral SDMT, the two WAIS-III speeded graphomotor subtests (DSy and SS) and the TMT. All three STROOP measures loaded on the third component (ST) which does not immediately suggest a generic description as it combines the three STROOP scores but does not load with recognised measures of verbal fluency or processing speed. The three subtests (DSp, AR, LNS) of the WAIS-III working memory index comprised the fourth component (WM). Finally, the fifth component provided a measure traditionally associated with verbal fluency (VF) comprised of COWAT and ANIMALS.

The status of the third component demonstrated the challenge and arbitrariness of characterising components in this type of research and can be considered in the context of the component intercorrelations presented in Table 6.3.

Table 6. 3

Component Correlation Matrix

Component	PS	ST	WM	VF
WK	.112	.226	.382	.284
PS		.424	.264	.250
ST			.254	.235
WM				.249

As indicated in the table, inferring that this component measures verbal fluency is belied by the fact that it correlates only .235 with component five (VF), which consists of measures that are commonly associated with the construct of verbal fluency. Similarly, the ST component correlates highest with the PS component (.424) which contains both verbal and visual measures. It is important to

note that while this is the highest intercorrelation, the strength of association was clearly insufficient to result in the Stroop measures loading on the PS component.

There is also an important lesson here for clinicians in their consideration of how tests map on to cognitive domains. Both the Trail Making Test and Stroop measures have variants that appear on the Delis-Kaplan Executive Functioning System (D-KEFS; Delis, Kaplan & Kramer, 2001). Understandably, clinicians may assume that this association would justify considering these two tests as measures of the same underlying construct, Executive Functioning. To the contrary, the PCA indicates that the Stroop measures are sufficiently different from measures of processing speed and verbal fluency, at least in this normative sample, to justify their interpretation as a separate construct no matter how superficially similar they may appear. It is humbling to reflect on how often in clinical research the roles and relatedness of measures are assumed rather than empirically verified (e.g. immediate and delayed memory composites of the WMS-III; Millis, Malina, Bowers & Ricker, 1999; Price, Tulsky, Millis & Weiss, 2002; Tulsky, Ivnik, Price & Wilkins, 2003; Tulsky & Price, 2003).

Based on the results of empirical investigation, measures may be selected to form valid composites. RAPT composite methodology may be used to incorporate this structure into test combination, analysis and interpretation. For example, an individual may be administered the COWAT and Animal Naming to examine functioning in verbal fluency. Typically, without the results of PCA or composite methodology this investigation may rely on examination of these individual test scores without evidence of their status as measures of this construct and regardless of their disparate reliabilities. Using composite methodology, however, these tests may be combined based on an empirical structure and with due reference to their shared variance (based on intercorrelation), individual levels of stability (based on test

reliability) and the stability associated with their combined interpretation (composite reliability). As indicated above the PCA structure would support the combining of COWAT and ANIMALS but would not support the inclusion of Stroop measures into such a composite.

For example, in a case in which the individual obtained a scaled score of 5 for the COWAT and of 6 for ANIMALS (transformed according to relevant normative samples using formula 5.1a), a verbal fluency composite may be obtained using composite methodology. Table 6.4 below lists the necessary normative information.

Table 6. 4

Normative Data for COWAT and ANIMALS

	Mean	**SD**	$\mathbf{r_{kk}}$	$\mathbf{r_{k(k-1)}}$
COWAT	10	3	.82*	.402***
Animals	10	3	.55**	

* COWAT reliability based on normative cases from the PCA sample.
** ANIMALS reliability based on normative cases from the PCA sample.
***Intercorrelation based on normative cases from the PCA sample.

Information from the table above is used to calculate:

- a composite sum of scaled scores of 11 (using formula 5.2),
- a composite mean of 20 (using formula 5.3a),
- a composite standard deviation of 5.024 (using formula 5.5),
- composite internal consistency coefficient of .775 (using formula 5.6),
- standard error of estimate of 6.264 (using formula 5.9a),
- and standard error of prediction of 9.480 (using formula 5.12a).

This information is then used to calculate:

- a Verbal Fluency deviation quotient of 73 (using formula 5.8a) reflecting a performance level associated with four percent of the population based upon the normal distribution;
- a 90 percent confidence interval of 69 to 89 (using formula 5.10); and
- a 90 percent re-test confidence interval of 64 to 95 (using formula 5.10a).

The detailed computations for this example are provided in Appendix A.

Using composite methodology clinicians could conclude that this individual's ability to process verbal information efficiently is estimated to fall within the Below Average range and is as good as or better than only four percent of the normal population. We can be confident with a margin of error of 10% that this individual's true score on this composite lies between 69 and 89. Further, if we were to administer these measures again we can be confident with a margin of error of 10% that this individual's verbal fluency score would fall within the range 64 to 95 based upon the psychometric properties of the composite. The implications of this is that re-test scores of less than 64 or more than 95 would signal a systematic change in behaviour.

This demonstration highlights the capacity of composite methodology as a direct means of applying empirically-based structure directly to clinical decision making. However, the implementation of RAPT methodology here highlights a number of issues that must be uppermost in the mind of the interpreting clinician. The computed confidence bands may seem large, but importantly this is the real-world consequence of using measures that generate a composite with internal consistency of only .775. Without the computation of these error bands clinicians are all too vulnerable to giving equal weight to all measures in their battery. The

composite reliability of .775 is, in fact, lower than the reliability of the COWAT alone and would give the reflective clinician valuable information about whether or not he or she should be employing a category fluency measure with higher reliability to more effectively measure the underlying construct and to reduce error variance. In the situation where clinicians have no choice but to employ these measures, they are able to modulate their inferences using composite methodology in accordance with the reliability of the measures they employ. The ultimate lesson of this is that clinicians must employ composite computations critically with the aim of, at least, understanding the psychometric implications of their decisions.

Similarly, when empirically validated structure is not available clinicians may use composite methodology to investigate test combinations based on theoretical (rather than empirical) structures. For example, a well known conceptual framework for classifying and understanding cognitive abilities is the theory of fluid and crystallised intelligence (Carroll, 1993; Cattell & Horn, 1978; Horn, 1988, 1998; Horn & Cattell, 1966; Stankov & Horn, 1980). This distinction may provide valuable information to the clinician on age-related changes in cognitive functioning by contrasting the education-related activities associated with "crystallised" measures and the speeded and analytical tasks typically defined by this structure as "fluid" (Kaufman & Lichtenberger, 2002). The concepts of fluid and crystallised intelligence have been mapped onto the structure of the WAIS-III but have not been incorporated into the methods available to clinicians in the technical manual (Wechsler, 1997a). Consequently, clinicians wishing to evaluate WAIS-III performance using the fluid-crystallised framework have limited choices available to them.

Without the use of any computations, test scores for fluid and crystallised intelligence may be analysed and interpreted individually using WAIS-III normative

data. Analysis of domain-based measures individually in this fashion, however, in effect assumes that all tests are equally reliable and equally valid as measures of the relevant constructs. This reasoning is psychometrically flawed and also provides little evidence about the shared relationship of tests as measures of crystallised and fluid abilities. In contrast, however, knowing the observed score, normative mean, standard deviation and reliability of each subtest and intercorrelations between subtests (available for all WAIS-III subtests), fluid and crystallised composites may be calculated facilitating both normative and ipsative analyses.

Of course, if only WAIS-III measures are employed in these investigations then norms-based adjustments from the WAIS-III standardisation sample, are available (Kaufman & Lichtenberger, 2002). However, if the clinician wishes to evaluate these domains using other tests or is unable to access adjustments, composite algorithms allow this structure to be formally applied providing scores which are at least as interpretable as those calculated using norm-based adjustments derived from the WAIS-III standardisation sample (as demonstrated in subsequent chapters).

In fact, the preceding demonstrations illustrate that composite methodology facilitates use of any validated theory of cognitive functioning as a means to structure selection, analysis and interpretation of any test battery. Additionally, the astute clinician would use composite methodology not only to formalise structure but to investigate the psychometric consequences of clinical decision making with the aim of understanding and reducing error. The capacity to facilitate such investigations constitutes the first fundamental strength of the RAPT method.

6.2.2 Composites Increase Domain-Based Reliability

The reliability of a composite may be readily calculated. This allows domain-based inferences to be drawn with a known degree of internal consistency. Knowledge of the degree to which random error is likely to be incorporated into clinical inferences based on either individual or composite test scores guides clinicians to an appropriate level of confidence using practical indicators of reliability, such as standard error.

Further, and perhaps more importantly, combining tests into composite scores may improve overall reliability, permitting a more stable measure of an underlying construct than any of the individual tests. According to classical test theory, increased sampling decreases the impact of random error and increases the influence of true scores. This suggests that a composite may be more psychometrically reliable and consequently less vulnerable to the impact of random error, than its component subtests. This is especially the case for subtests with lower reliability.

As demonstrated in the preceding example, however, this is not always the case: the combination of a highly reliable subtest with a subtest of low reliability in the VF component of the PCA produced a composite which was less reliable overall than the more stable individual test (COWAT). However, this example does not disprove the capacity of composites to capitalise on the increased stability likely with multiple domain sampling and instead alerts clinicians to a point which is intentionally belaboured in the current thesis that the process of considering the influence of reliability, and psychometric characteristics in general, cannot be left to clinicians' assumptions or longstanding heuristics. While statistically speaking increased sampling may diminish the magnitude of random errors associated with measurement of a particular cognitive construct, practically the clinician must seek concrete evidence of the reliability of actual test combinations. In many instances

evidence regarding this reliability may be gathered using composite methodologies. In the VF component, if the goal is merely the most accurate measurement of verbal fluency, then the use of COWAT alone would be psychometrically justified. If, however, measures of both letter and category fluency were specifically wanted then the results of both the PCA and composite reliability would strongly suggest either the selection of a more reliable category fluency task or the inclusion of additional category trials. For example, the COWAT score consists of three 60-second trials, while ANIMALS consists of a single 60-second trial.

To investigate the heuristic provided above, that increased sampling should decrease the impact of random error, simulated composite reliabilities were computed, using formula 5.6, based on subtests with varying levels of internal consistency and moderate intercorrelations (i.e., .5). Specifically, composite reliability coefficients were based on combinations of two subtests with equal reliabilities ranging from .5 to .9. Each simulated composite was based on a combination of two tests with the same reliability. For example, the first row of Table 6.5 demonstrates the composite reliability associated with combining subtests with coefficients of .5. This fairly unrealistic situation was constrained specifically to evaluate whether (a) reliability improved with the addition of subsequent subtests, and (b) whether this phenomenon varied when subtests were moderately to highly reliable, without the confound of differing subtest reliabilities in the composite. All subtest intercorrelations were constrained to .5: a moderate level of interrelationship between tests. Again, this was done with the specific aim of isolating the consequence, in terms of reliability, of forming composites without the confound of varying interrelationships.

Table 6.5 compares the reliability coefficients of these simulated composites, the calculations for which are included in Appendix B. The results of the simulation

indicate that combining two subtests with equal reliability and moderate interrelationship increases the reliability of measurement overall. For example, if two subtests with a reliability of .5 are combined the resulting composite obtains a reliability coefficient of .67. However, the combination of two highly reliable measures (each of .9) produces only a modest increase in composite reliability (r = .93). Thus the improvement in composite reliability is less marked as subtest scores increase in reliability. Combining tests of the same reliability, however, leads to a reduction in the vulnerability of measurement to random error: a distinct advantage of composite methodology.

Table 6. 5

Comparison of Individual and Composite Reliability

Individual Test	**2 Subtest Composite**
r = .50	r = .67
r = .60	r = .73
r = .70	r = .80
r = .80	r = .87
r = .90	r = .93

As demonstrated in the previous section, however, not every combination of subtests is guaranteed to increase reliability. Composite reliability will be moderated by the reliability of all component tests and in some instances individual subtests will be more reliable than the composite overall. However, the simulation suggests that an advantage of combining tests scores of the same domain into composite scores is potential improvement in the overall reliability of the resulting composite. This is particularly likely when subtests are similarly vulnerable to the impact of random

error or have weak reliabilities. Results suggest that clinicians, who must of necessity use measures with low internal consistency, could offset the negative impact by increasing the number of measures and combining them into a composite score. The potential for improving the stability of domain-based test scores is a second strength of the RAPT methodology.

6.2.3 Composite Reliability Guides Test Selection

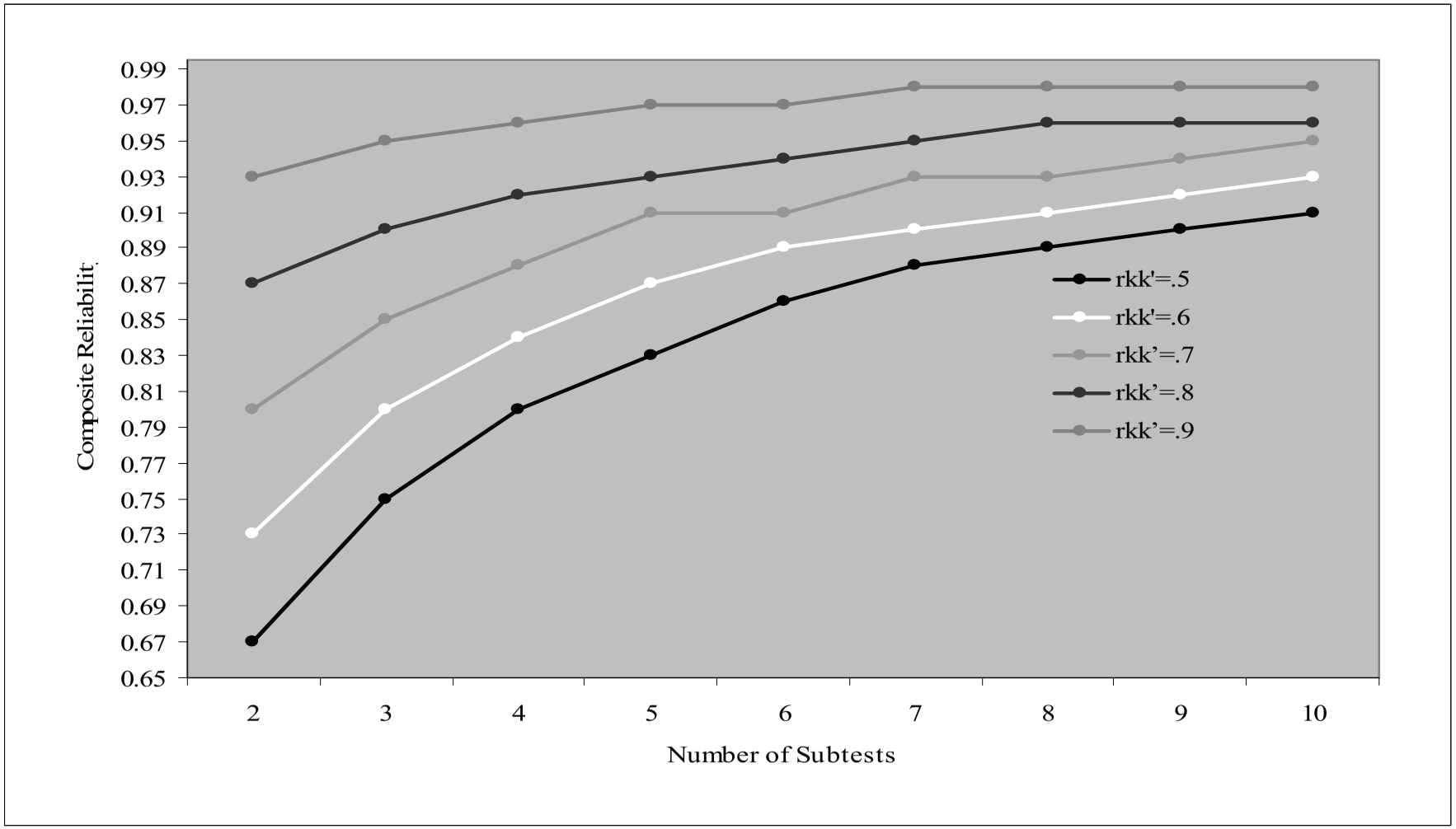
Reducing the number of tests administered is desirable for practical reasons, such as accommodating time constraints and decreasing the impact of fatigue, as well as for statistical reasons. Increasing the number of tests increases the chance of finding impairment simply due to chance and this can be offset by combining tests into a composite where a single score with potentially greater reliability is the focus of interpretation. In this way, evaluation of the reliability of a composite can facilitate testing efficiency as the minimum number of tests required to measure a construct to the maximum degree of reliability is known.

While construction of a composite generally improves reliability above that of the individual subtest components, adding subtests beyond a specific point results in only small improvements in reliability. When measuring cognitive domains, clinicians may be administering more measures than necessary insofar as reliability is not improved. The reliability associated with domain based inferences can dictate both the number and type of tests selected and hence provide a guide to battery construction based on psychometric principles.

To investigate this potential strength, simulated composite reliabilities were again calculated based on combination of subtests with equal reliabilities again ranging from.5 to .9 with intercorrelations held constant at .5 (see Appendix C for a table of composite reliabilities). The objective of the current simulation was to

determine whether composite reliability coefficients failed to increase beyond a certain point. To this end, two to ten subtests composites were calculated based on equally reliable subtests with coefficients of .5, .6, .7, .8 and .9. Again, the aim was to isolate the consequence of adding subtests in terms of reliability without the confound of either test inter-relationships or variations within subtest reliabilities. The results of this simulation are discussed below and depicted in Figure 6.1.

Figure 6.1 Improvements in composite reliability coefficients as subtests are added



For example, a two subtest composite with component subtests of equal and high reliability ($r_{kk'}$ = .9) and moderate intercorrelation (.5) had an overall composite reliability of .93. Addition of another, equally reliable subtest improved the overall composite reliability by only .02 to $r_{kk'}$ = .95. If a similar composite was calculated with subtests of a much lower reliability, such as $r_{kk'}$ = .5, and the same moderate intercorrelation of .5, the addition of a subsequent and equally reliable subtest improved the overall composite reliability to a much greater degree, from an index of .67 to .75. In the first example adding subsequent tests did not result in a substantial improvement in reliability and hence may not be warranted in terms of the additional effort, time and risk of artifactual error due to over-sampling. In the second example, the addition of even further tests may be warranted given the improvement in the overall reliability of the composite. In this way composite reliability can be used to evaluate the role each test plays in the battery as well as to support more stable measurement of the particular construct of interest even when clinicians use measures with only modest reliabilities.

In practical terms, a clinician who must use measures that have reliability less than the oft-quoted standard of .8 may manage potential unreliability by using more of these measures and combining them into a composite. The resulting composite score is like to have greater reliability, without the risk of an inflated type I error resulting from administering increasing numbers of measures and attempting to interpret each measure separately.

Overall, the advantage, in terms of composite reliability, of adding subtests diminishes as subtest reliabilities increase. For all composites the addition of a fifth subtest failed to substantially increase composite reliability. However, for composites with particularly high subtest reliabilities (.8 – .9) composite reliability

increased little beyond that which was achieved with the combination of as few as three subtests.

This phenomenon is perhaps most emphatically demonstrated by evaluating the percentage of improvement in reliability gained by adding additional subtests to two subtest composites. As with all preceding examples, subtest reliabilities are equivalent and intercorrelations between measures are constrained at .5 for consistency (See Table 6.6).

Table 6. 6

Percentage of Improvement in Composite Reliability when Subtests are Added to a Two-Subtest Composite

Subtests in Composite	**Subtest Reliability**				
	$r_{kk'} = .50$	$r_{kk'} = .60$	$r_{kk'} = .70$	$r_{kk'} = .80$	$r_{kk'} = .90$
3	8%	7%	5%	3%	2%
4	5%	4%	3%	2%	1%
5	3%	3%	3%	1%	1%
6	3%	2%	0%	1%	0%
7	2%	1%	2%	1%	1%
8	1%	1%	0%	1%	0%
9	1%	1%	1%	0%	0%
10	1%	1%	1%	0%	0%

The reliability coefficients of all five composites improve with the addition of an extra subtest (i.e., to form a three subtest composite), with the greatest improvement (i.e., 8%) obtained by the composite with the least reliable component subtests. Addition of a third subtest, to form a four subtest composite, also improved

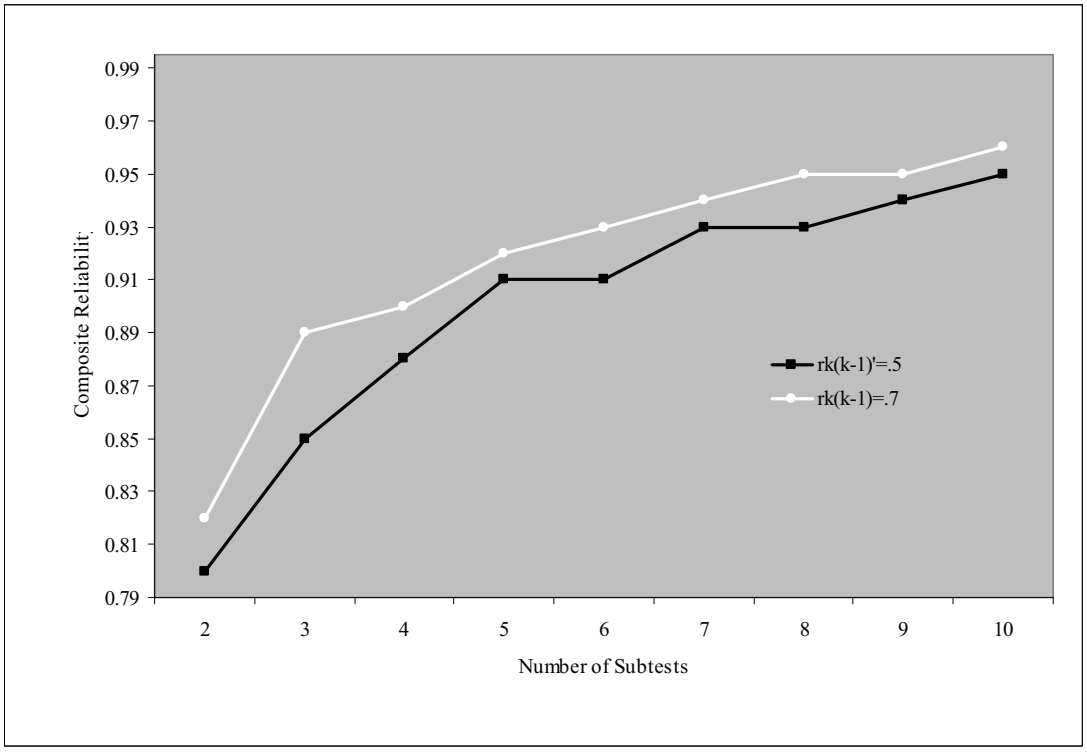
the overall composite reliability for composites with component subtests of low ($r_{kk'}$ = .5) and moderate ($r_{kk'}$ = .6 - .7) reliability. This improvement was minimal, however, for composites with highly reliable component subtests ($r_{kk'}$ = .8 – .9). Little improvement in composite reliability was achieved by calculating composites with six or more subtests, regardless of the reliability of component subtests. The results of this demonstration suggest that when subtest reliability is at least moderate (i.e., equal to or greater than .6) to reliably measure a well defined construct five subtests are sufficient: few gains in composite reliability were achieved in the simulated example by the combination of six or more subtests of a particular domain unless subtests were moderately unreliable. As previously discussed, both individual subtest reliabilities and subtest intercorrelations also exert influence on the reliability of a subsequent composite. While the current example provides an argument for the use of reliability in guiding domain-based test selection, it must be moderated in clinical practice by direct examination of the actual reliability associated with clinical composites. Again, the current demonstration cannot dictate a heuristic for test selection but rather provide clinicians with a rationale for considering the consequences of their proposed battery. The potential to guide battery selection based on a priori examination of the reliability of proposed composites constitutes the third strength of the RAPT model.

6.2.4 Composites Facilitate Valid Test Combinations

The tendency of composite reliability to increase with the incorporation of additional tests is even more pronounced as the tests increase in validity. The following example extends the simulated composite reliabilities calculated in section 6.2.3 by calculating similar composite reliabilities based on tests with stronger

intercorrelations ($r_{k(k-1)}$ = .7). In this instance, the impact of subtest interrelationship on reliability is the focus of interest and is directly evaluated by comparison of composites with moderate (.5) and strong (.7) intercorrelations. Figure 6.2 compares the reliabilities of composites composed of subtests with intercorrelations of .7, with those comprised of tests with more moderate test intercorrelations, of .5. For both composites, the reliabilities of components tests are moderate (i.e., $r_{kk'}$ = .7). In this instance, subtest reliability coefficients are constrained to avoid their demonstrated impact on composite reliability and to specifically isolate the consequences of increased inter-relationship between component tests. As in the previous section, composites are based on combinations of two to ten subtests to identify trends associated with the number of subtests. A table of coefficients are provided in Appendix D.

Figure 6.2 Composite reliability with tests of moderate (r = .5) and high (r = .7) intercorrelations, when reliability of tests is constrained at .7



As demonstrated in figure 6.2, the reliabilities of composites composed of tests with high intercorrelations ($r_{k(k-1)}$ = .7) demonstrate the same general increases as composites composed of subtests which are less strongly related ($r_{k(k-1)}$ = .5). However, tests which share stronger intercorrelations produce composites which are more reliable overall than composites comprised of equally reliable tests which share weaker intercorrelations. In other words, the validity of the composite, in terms of test intercorrelations, improves the stability of domain measurement overall. While at the level of the individual test, reliability provides an upper limit to validity, at the composite level validity facilitates improved reliability. This trend holds true for composites composed of tests with higher reliability ($r_{kk'}$ = .8; Figure 6.3) though decreases as test reliabilities near 1 ($r_{kk'}$ = .9; Figure 6.4).

Figure 6.3 Composite reliability with tests of moderate (r = .5) and high (r = .7) intercorrelations, when reliability of tests is constrained at .8

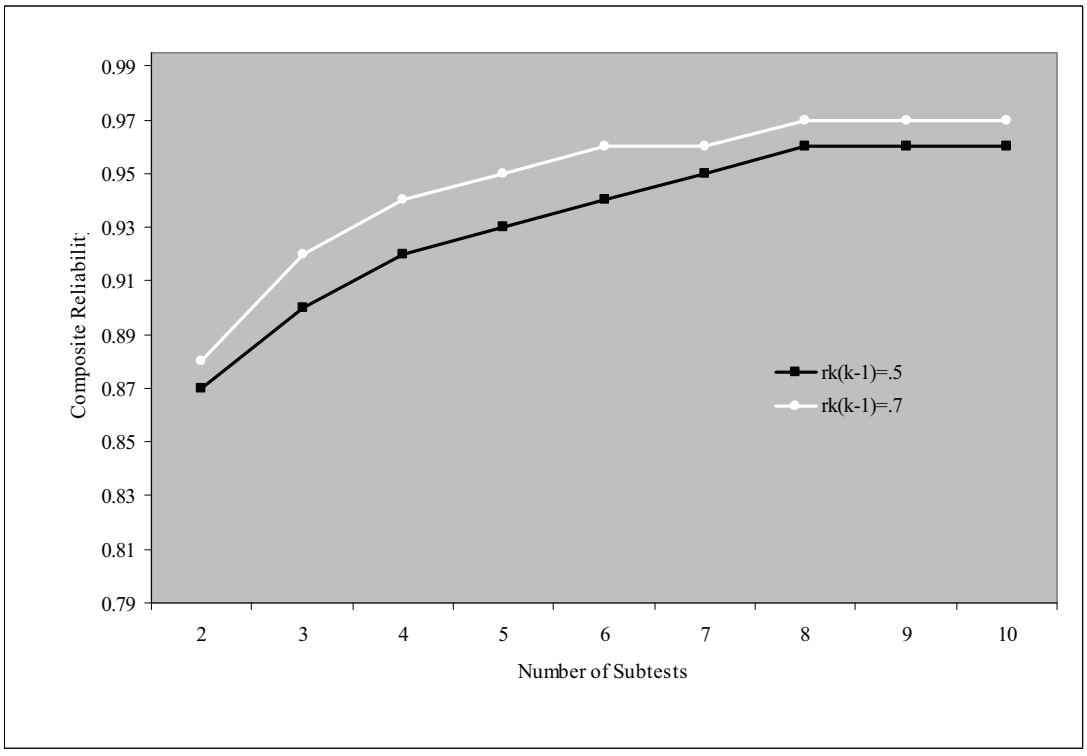
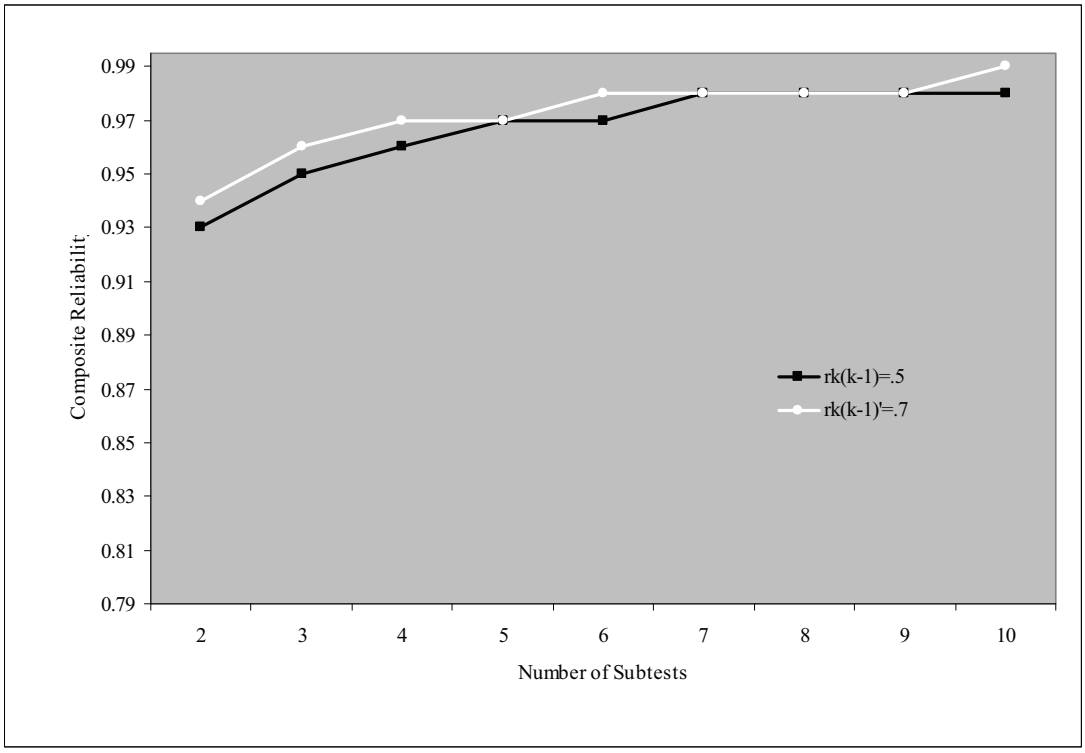


Figure 6.4 Composite reliability with tests of moderate (r = .5) and high (r = .7) intercorrelations, when reliability of tests is constrained at .9



As indicated in figures 6.3 and 6.4, composites with more strongly related subtests tend to have reliabilities which surpass those comprised of equally reliable subtests with more modest intercorrelations. However, as subtests become increasingly reliable it is this reliability, rather than the strength of their intercorrelations, which dictates the overall reliability of the composite. Specifically when subtests are highly reliable, as in figure 6.4, improvements in intercorrelations produces only a small and somewhat inconsistent improvement in composite reliability.

This trend is most succinctly demonstrated through a direct comparison of the degree that composite reliability coefficients increase as subtest intercorrelations increase from .5 to .7. Specifically, in this examination, the improvements gained through increasing subtest intercorrelations were examined when subtest reliabilities were constrained at .5 (Composite Discrepancy 1), .6 (Composite Discrepancy 2), .7 (Composite Discrepancy 3), .8 (Composite Discrepancy 4) and .9 (Composite

Discrepancy 5), for two to ten subtest composites. Table 6.7 outlines the results of subtracting reliability coefficients, based on strongly related subtests, from those based on subtests sharing modest correlations.

Table 6. 7

Improvement in Composite Reliability Coefficients when Subtest Intercorrelations are Increased from .5 to .7

Composite Discrepancy	**1**	**2**	**3**	**4**	**5**
2 subtests	.04	.01	.02	.01	.01
3 subtests	.04	.03	.04	.02	.01
4 subtests	.04	.03	.02	.02	.01
5 subtests	.04	.02	.01	.02	.00
6 subtests	.03	.02	.02	.02	.01
7 subtests	.02	.02	.01	.01	.00
8 subtests	.03	.02	.02	.01	.00
9 subtests	.02	.02	.01	.01	.00
10 subtests	.02	.02	.01	.01	.01
Average Discrepancy	.03	.02	.02	.01	.01

The reliability coefficient associated with a two test composite with subtests of equally modest reliability (.5) improved by four points when test intercorrelations increased from .5 to .7 (Column 1 of Table 6.7). In this instance, the reliability coefficient increased from .67 to .71 when the relationship between subtests strengthened and all other factors were held constant (see Tables C.1 and D.1). Examination of discrepancies further down this column indicate similar

improvements when subtest reliabilities were low (.5) and, in columns 2 and 3, demonstrable but less consistent improvements when subtest reliabilities were moderate (.6, Column 2; .7, Column 3). When subtest reliabilities were strong, however, improving the relationship between component tests has little impact on composite reliability (Columns 4 and 5).

Overall, the greatest improvements were gained when subtest reliabilities were low ($r_{kk'}$ = .5 – .7). In these instances the average discrepancy between composite reliabilities ranged from .3 to .2 coefficient points. As the reliability of component tests increased, however, their interrelationships impacted less on the overall stability of the composite which instead was increasingly a function of subtest reliabilities. For example, the reliability coefficients of composites with high subtest reliabilities (.8, Column 4; .9, Column 5) improved very little (an average of only .01 coefficient points) when subtest intercorrelations increased from .5 to .7. Regardless, validity undoubtedly plays an important role in improving the overall stability of domain-based measurement. This simulation demonstrates that this may particularly occur when component tests have only modest reliabilities.

At this stage, it is notable, that this demonstration is largely for illustrative purposes as tests with internal consistency coefficients of .5 or .6 should not be able to correlate at .7. Validity cannot, theoretically speaking, exceed reliability and where this does occur it can be suspected that reliability is underestimated. Despite this psychometric truth, however, the current example highlights the capacity of composite methodology to reflect both the reliability and the validity associated with domain-based inferences. As with previous demonstrations, proposing rules-of-thumb for battery selection is not an aim of investigation which instead is undertaken to detail the capacity of composite scores to reflect psychometric characteristics. In

the current example, the fact that composites reflect not only subtest reliability, but subtest relationships, constitutes another potential advantage of using RAPT methodology as a rationale for flexible battery selection and analysis.

6.2.5 Composites Moderate Interpretative Errors

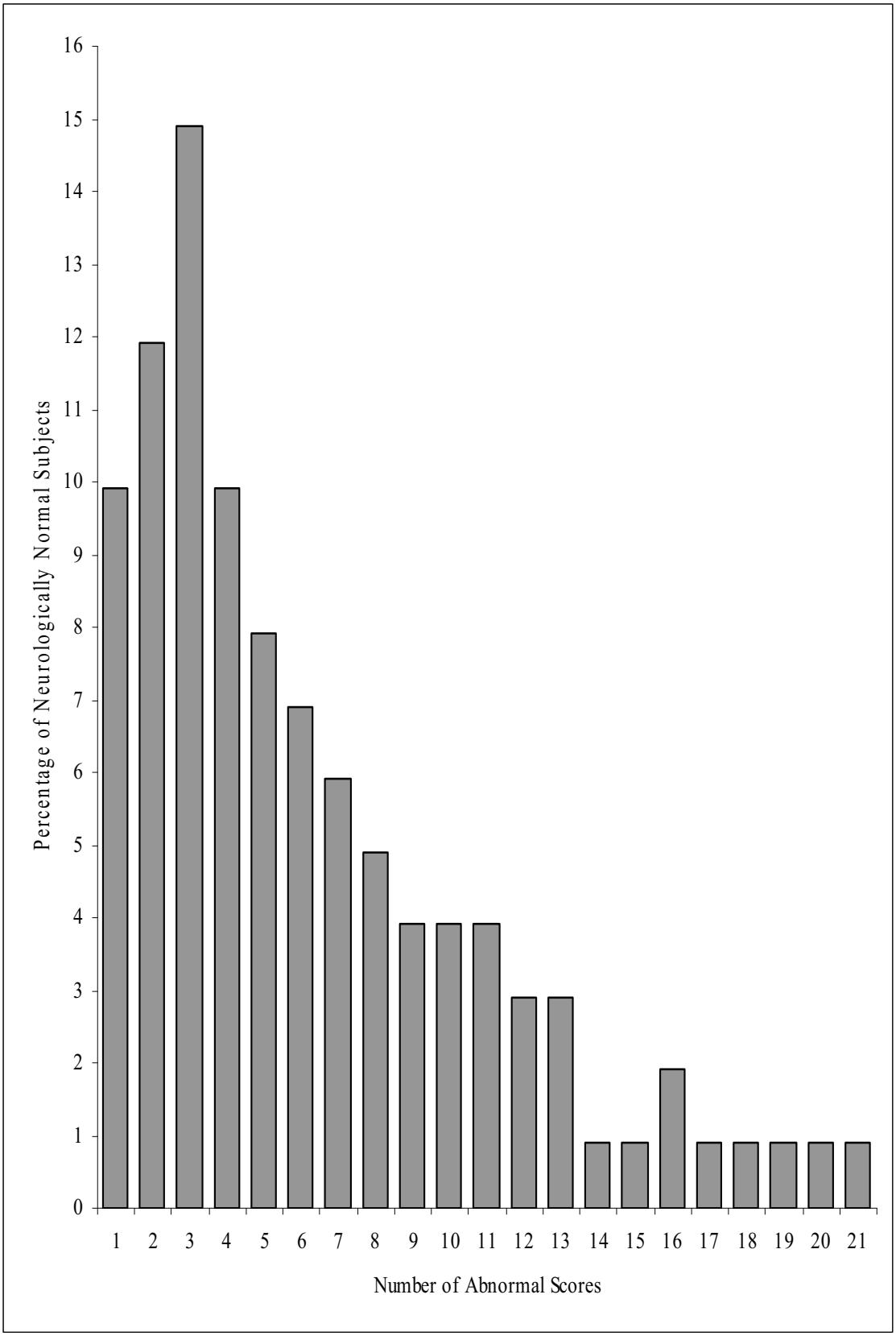
The robustness of composite scores as a measure of specific cognitive domains also provides a rationale for usage, as composite methodology reduces the impact of artifactual errors to which individual tests are highly vulnerable. When interpreting individual test scores a single "impaired" score may be clearly identified. Such a score may, however, occur without the presence of true impairment. For example, if a clinician utilises a frequency of occurrence of 10% as a criterion for detecting the presence of impairment, logically this would result in a false alarm rate amongst the normal population of 1 in 10 or 10%. It is important to note that the use of 10% is entirely arbitrary and that whatever criterion is used a requisite proportion of the normal population will be falsely identified. Using more stringent criteria such as 5% or 1% frequency of occurrence merely reduces the number of false alarms in the normal population and increases the number of misses in the clinical population. As previously stated, test scores may appear to be "impaired" for a variety of reasons unrelated to true impairment in the construct of interest and individual impaired scores may be obtained despite average or above average overall performance. Ingraham and Aikken (1996) demonstrated that the likelihood of obtaining at least one score in the impaired range increases as more tests are administered.

The phenomenon of normally functioning individuals obtaining one or more scores within an "impaired" range can be demonstrated on several major batteries. Heaton, Grant and Matthews (1991) for example state that "some poor test results are

to be expected in most normal persons, especially when a large battery of tests is administered" (p.36). In demonstrating this, they provided results from a sample of 455 neurologically normal individuals administered forty measures of the WAIS and HRNB which emphatically demonstrated the prevalence of impaired scores (T score ≤39) in "normals". The results of this analysis are reproduced in Figure 6.5 below and indicate that abnormally low scores occurred in marked proportions. Remarkably, ninety percent (90%) of the "normal" sample obtained "abnormal" scores (T ≤39) on at least one battery subtest with a small proportion obtaining up to twenty-one impaired scores (Heaton, Grant & Matthews, 1991, p.36-37).

Figure 6.5 emphatically illustrates the vulnerability of individual scores on this battery to artifactual error. In fact in interpreting individual HRNB scores the clinician is likely to risk type I errors in a vast majority of cases. There is some evidence, however, that combining individual scores into composites protects against such inflated error rates. Rojas and Bennet (1995), for example, demonstrated that HRNB composite scores (i.e., General Neuropsychological Deficit Scale, GNDS, and the Halstead Impairment Index, HII) were far less vulnerable to such artifactual errors. These authors demonstrated that both the GNDS and HII outperformed the ability of any single HRNB measure in accurately discriminating between normal test takers and those with mild brain-impairment. This research arguably demonstrates that clinicians could use HRNB composite scores with greater confidence in their ability to draw correct clinical inferences than that warranted by use of any single HRNB test.

Figure 6. 5 Percentage of impaired subtest scores in the CNEHRB normative sample



(Derived from Heaton, Grant & Matthews, 1991)

This trend is also apparent when cognitively "normal" individuals are examined using subtests of the WAIS-III and WMS-III. As indicated in Table 6.8 below, in the WAIS-III/WMS-III co standardisation sample (n=1250), almost half of the sample obtained one or more scores which were at less than the tenth percentile (45.5% for the WAIS-III, 50.2% for the WMS-III) with a smaller percentage obtaining scores lower than the second percentile (15.2% for the WAIS-III, 22.0% for the WMS-III) (Olm & Senior, 2006).

Table 6. 8

Comparison of the Percentage of Normal Individuals Obtaining Impaired Scores on WAIS-III and WMS-III Subtests and Composites

WAIS-III	**<10th %ile**	**<2nd %ile**
Subtests	45.5	15.2
Composites	17.6	4.4
WMS-III	**<10th %ile**	**<2nd %ile**
Subtests	50.2	22.0
Composites	24.4	7.0

As demonstrated in the table, use of composite scores substantially reduced the proportion of individuals in the normal sample who may have been misclassified based on single impaired subtest scores. For example, while for the WAIS-III approximately fifteen percent (15.2%) of normal individuals obtained a subtest score indicative of severe impairment (i.e., less than the 2nd percentile), fewer than five percent (4.4%) obtained a WAIS-III composite score of lower than the second percentile.

These examples illustrate the capacity of composite scores, at least on the HRNB, WAIS-III and WMS-III, to provide some protection against errors of misclassification due to artifactual errors. Interpreting individual test scores on these batteries without consideration of their intrinsic interrelationships may increase the chances of erroneous clinical inferences. The RAPT methodology applies this advantage to domain-based composites comprised of any individual tests which may validly be combined.

6.2.6 Composites Facilitate Flexibility in Test Selection

The previous examples provide specific demonstrations of the psychometric and statistical strengths of composite methodology. Most importantly using composite scores provides a means of achieving psychometric robustness while paying due attention to the competing demands of clinical practice. RAPT is intended as a fundamentally flexible methodology in which the "toolbox" is fixed but the method of employment is entirely a function of the clinical demands and issues relevant to unique clinical situations. A fundamentally important strength of composites is their capacity to provide substantial robustness to the interpretation of test scores regardless of the flexible battery structure often necessitated by the rigours of clinical practice.

As has been discussed in the preceding sections the strength of the demonstrable psychometric characteristics of composites is not in their capacity to formalise heuristics of test selection and combination, but rather in their capacity to facilitate answers to psychometrically based questions which arise as clinicians select and use testing batteries. Reconsidering the PCA presented earlier in this chapter in

terms of this highlights the implications of the RAPT methodology with regard to test selection, analysis and interpretation.

While the reliability of composites provides a rationale for test selection, the clinician must consider issues of validity when choosing tests of a particular domain. Component one in the PCA presented earlier in this chapter consists of seven measures of verbal ability or word knowledge. In terms of a testing battery, this component may be being over-assessed with no substantial improvement in reliability. In the absence of composites, interpretations based on validity information alone would dictate the interpretation of all seven measures with the aim of drawing inferences about word knowledge. This would, however, increase the probability of a type I error given the higher likelihood of one of these measures scoring outside of the average range simply due to chance. On the other hand, using composite methodology the clinician may evaluate the reliability of potential test combinations to maximise reliability of the resulting composite.

For example, the tests which load on WK in the PCA did not contribute equally to this component and instead may be ranked based on loadings as indicated in Table 6.9, which also includes the reliability coefficients for each measure. As indicated in this table, some tests are stronger measures of word knowledge than others and may be chosen for a word knowledge composite based on their relative loadings. For example, WRAT-3 Reading or STW may be omitted from the composite based on weaker loadings. Alternately the subtests which load most highly on the component, Information, BNT and Vocabulary may be combined to produce a composite highly representative of the underlying, empirically determined component.

Table 6. 9

Word Knowledge Subtests Ranked According to Component Loadings

Test	Component Loading	Reliability Coefficient
WAIS3-IN	.852	.91*
BNT	.804	.81**
WAIS3-VO	.801	.93*
WAIS3-CO	.728	.84*
WAIS3-SI	.711	.86*
STW	.655	.83***
WRAT3-Reading	.650	.95****

* IN, VO, CO and SI reliability coefficients based on Wechsler (1997a).
** BNT reliability coefficient based on normative cases from the PCA sample.
*** STW reliability coefficient based on Baddeley, Emslie & Nimmo-Smith (1992).
****WRAT-3 Reading reliability coefficient based on Wilkinson (1993).

The degree, however, to which each test is a valid measure of the construct, must be moderated in terms of reliability. As indicated in the third column of the table, if tests were ranked according to reliability a very different arrangement would be obtained. If test selection were based solely on individual reliability, clinicians may omit the BNT or Comprehension as the least reliable measures and instead combine the three most reliable subtests, Reading, Vocabulary and Information. Such a composite clearly differs from that proposed through consideration of component loadings alone.

Moderation of this conflict may be resolved through consideration of clinical value. For example, while the BNT is the least reliable measure on the WK component and could be omitted from a word knowledge composite on these grounds, it has the second strongest loading and the additional advantage of being highly sensitive to the influence of a variety of pathologies (Williams et al., 2007). The clinician may, therefore, choose to include such a measure for its dual purpose

and strong loading on the WK domain. Similarly, WRAT-3 Reading has the weakest loading on the construct of Work Knowledge, however, is the most reliable subtest and also provides the advantage of providing additional information regarding grade-based reading level which may serve an important role in consideration of consistency with school records or informing the clinician with regards to the most appropriate psychosocial inventory to employ. It may be included both for dual purpose and for its strong contribution to the empirically determined domain. Finally, given that STW is the only measure of verbal ability which does not rely on actual verbal output, clinicians may choose their measure to elicit a client's word knowledge without the requirement of verbal output.

Clearly the decisions relating to test selection for this single domain are numerous and difficult to resolve and clinicians may be sympathised with for avoiding them. However, RAPT provides a clear methodology for resolving just such confusion as the overall reliability of various potential subtest combinations may be readily calculated and considered. Table 6.10, for example, provides the reliability coefficients of the potential composite combinations discussed above. It is clear that composite reliability provides vital information in resolution of the above dilemmas (see Appendix E for composite reliability computations).

Table 6. 10

Composite Reliability Coefficients for Several Verbal Composites

Rationale	Tests	Composite Reliability
Most valid	IN, BNT, VO	.95
Most reliable	Reading, VO, IN	.97
Most practical	Reading, STW, BNT	.93

As indicated in the table, composites may be formed based on several conflicting rationales, however, consideration of composite reliability allows the clinician to make an informed choice between several seemingly viable options. As expected, the composite comprised of the most highly reliable subtests is the most reliable composite. However, the most practical test combination still achieves a high level of reliability and could provide an alternate which is valid, reliable and clinically useful. These different concerns are at the very heart of semi-flexible batteries in that clinicians need to be able to alter their test batteries to best suit the needs of the assessment and their clients. In the absence of a formalised means of evaluating reliability at the battery level, however, clinicians are denied the necessary information with which to offset the consequences of their decisions in terms of measurement error. Perhaps more importantly than its role in test selection, however, is the capacity of RAPT methodology to allow indices of domain-based error to be directly incorporated into subsequent analyses and interpretations.

Even when clinicians do not choose the most reliable combination, the error associated with their choice will be known and incorporated into analyses and interpretations. In the current example, standard errors are directly incorporated into the inferential framework. The fundamental point is that the clinician does not simply have to choose tests based on habit, availability, individual reliability or various disparate or conflicting validity studies, but may evaluate the degree of confidence warranted by any test combination. The fact that RAPT facilitates use of the reliability of the composite score to empirically guide test selection regardless of which tests are chosen, is its greatest strength.

6.3. Conclusions

The examples above demonstrate the capacity of composites to improve the psychometric qualities of a test battery. Few clinicians would, however, give psychometric factors precedence over clinical or practical considerations. To the contrary, flexible application of composite methodology ensures that both psychometric and clinical factors can be adequately considered by providing a systematic means of evaluating the psychometric consequences of necessary clinical decisions, even when such decisions reduce the overall psychometric quality of the battery.

The composite methodology employed in RAPT should not lead to a convergence to a single battery to be used on all occasions. Undoubtedly such a battery could meet exacting psychometric standards and avoid potentially complex considerations at the battery level. A standardised battery is highly likely to occur at the expense, however, of clinical usefulness. In any particular assessment some considerations may outweigh others and clinicians need to modify their test batteries accordingly. The construction of composites recommended in RAPT methodology allows formalisation of a process of integration between the equally compelling considerations of reliability, validity and clinical flexibility. For example, an optimal battery under RAPT methodology in a clinical context where a maximum of two hours is available for testing will undoubtedly result in a different selection and combination of tests from an assessment in which four or eight hours is available for testing. Similarly, an assessment of an individual with a major sensory or motor impairment would demand different considerations than one in which all sensory and motor modalities are intact. In each instance, the composite methodology recommended in RAPT facilitates the consideration of the psychometric

consequences of all assessment decisions despite the ever-changing nature of the clinical battery.

Psychometric evaluation at each level of battery usage may provide contradictory or even opposing suggestions which must be accommodated for by flexibility of battery construction. For example, the methodology allows evaluation of the number of tests necessary to obtain optimal reliability. As indicated above almost no improvements, in terms of reliability, are gained by the combination of more than two tests of high reliability (i.e., .9) which have moderate intercorrelations. A clinician could not, however, base test selection on consideration of these psychometric characteristics alone and go on to argue that such a composite was not vulnerable to impaired scores (i.e., in this case a single impaired score would substantially influence the composite) or that the resulting composite score provided comprehensive coverage of a complex domain of functioning. In other words, choice of the most reliable composite may lead to insufficient coverage of a domain or increase the possibility of type I errors in inference. These considerations provide a pragmatic rationale, apart from consideration of reliability alone, for the choice of three tests. Alternatively, the clinician may choose a composite comprised of several tests which share strong intercorrelations yet with moderate reliabilities. The essential point is that in each instance using composite calculations the overall reliability of various combinations may be evaluated allowing the clinician to construct a battery which effectively balances reliability, validity and practical clinical utility. Possibly, it is this capacity which provides the greatest strength of the composite methodology proposed in RAPT.

CHAPTER SEVEN

REPLICATING KNOWN FIXED BATTERIES USING RAPT

7.1 Introduction

In the previous chapter, the use of RAPT methodology to integrate competing psychometric and clinical demands was demonstrated. In the current chapter, RAPT methodology will be applied to the replication of existing cognitive composites which demonstrate current reliability, validity and clinical applicability. Specifically, composites from the WAIS-III and Wide Range Achievement Test, Fourth Edition (WRAT-4) will be reproduced using RAPT methods and compared with analyses for the same measures outlined in their respective manuals. The intent of RAPT methodology is to provide the level of data analysis customarily available to, and where used a fundamental strength of, fixed batteries. Notably, RAPT methodology facilitates the type of clinical conclusions possible using existing systems such as the Wechsler interpretive methodology.

In fact, at this stage it would be appropriate to acknowledge the influence of Wechsler methodology in the development of the RAPT method. For example, the composite measures employed in “batteries” such as the WAIS-III, WMS-III or WISC-IV and CMS capitalise on the psychometric advantages of an unchanging collection of measures. Each of these groups of measures is based on cognitive structures which have been extensively empirically evaluated. They employ composite methodology with its attendant advantages, they allow the clinician direct consideration of measurement error through confidence intervals and provide analytical structures which facilitate the empirical evaluation of strengths and weaknesses (Wechsler, 1997a). The type of analysis common to the Wechsler scales served as a model for the method presented here and its validation is achieved, in no

small part, through demonstrating the capacity of RAPT methodologies to replicate the structure and analytic procedures employed within the Wechsler scales.

Essentially the current chapter examines how accurately the mathematical approaches presented in chapter five and evaluated in chapter six replicate the methods provided in existing, well known and widely accepted, measures. In perhaps more grandiose terms, if the RAPT methodology can be demonstrated to replicate the "known world" then clinicians will be more confident that it can effectively be employed in "the great unknown". If adequate accuracy is demonstrated then clinicians can apply RAPT methodologies to their own semi-flexible batteries with confidence in their ability to moderate the often competing demands of reliability, validity, and normative issues in battery usage.

To this end, this chapter will demonstrate the use of RAPT methodology in replicating the summary scores and various interpretive tools such as composite reliabilities, standard errors, confidence intervals, and, summary score intercorrelations, and percentiles of current widely used and accepted "batteries". First RAPT methodology will be used to replicate the Reading Composite of the Wide Range Achievement Test, Fourth Edition (WRAT-4) and comparison will be made between the resulting composite and the composite based on normative analyses provided in the professional manual for this test (Wilkinson & Robertson, 2006). Next, RAPT algorithms will be used to calculate Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) summary scores based on the normative sample. In this section investigation will focus on evaluating the capacity of RAPT methodology to replicate the WAIS-III analytical structures in terms of both the normative and ipsative comparisons available in the technical and administrative manuals and widely used by clinicians (Wechsler, 1997a, 1997b). It is hoped that

these demonstrations will provide clinicians with the degree of confidence in the capacity of the RAPT which would warrant its application beyond existing analytical frameworks to flexible collections of cognitive tests.

7.2 Replicating WRAT-4 Reading Composite

The WRAT-4 is the most recent edition of a battery designed to measure the "basic academic skills necessary for effective learning, communication, and thinking" (Wilkinson & Robertson, 2006, p.1) using subtests of reading, spelling and arithmetic. Unlike previous editions, which were analysed and interpreted using individual test scores (i.e., Wide Range Achievement Test, Third Edition, WRAT-3; Wilkinson, 1993) the revisions published in the WRAT-4 include a measure of sentence comprehension which is combined with the measure of word reading to produce a reading composite score. The following outlines the replication of scores and interpretive tools for this composite score using RAPT algorithms.

7.2.1 WRAT-4 Reading Composite Percentiles and Confidence Intervals

Formulae 5.2, 5.3, 5.4, 5.6, 5.8, 5.9, 5.10 and 5.11 were used to calculate an alternate Reading Composite (RC) to replicate that provided in the WRAT-4 manual (Wilkinson & Robertson, 2006). Median internal consistency coefficients for both the Blue (Word Reading $\alpha = .92$; Sentence Comprehension $\alpha = .93$) and Green forms (Word Reading $\alpha = .92$; Sentence Comprehension $\alpha = .93$) and a subtest intercorrelation for the age-based standardisation sample participants ranging in age from 19 to 94 years of $r = .68$ were used in calculations. A section of results from these analyses are compared with the normative indices provided in the professional manual tables in Table 7.1. This table compares the deviation quotients and ninety

percent confidence intervals calculated using RAPT methodology with those from the WRAT-4 manual.

Table 7.1

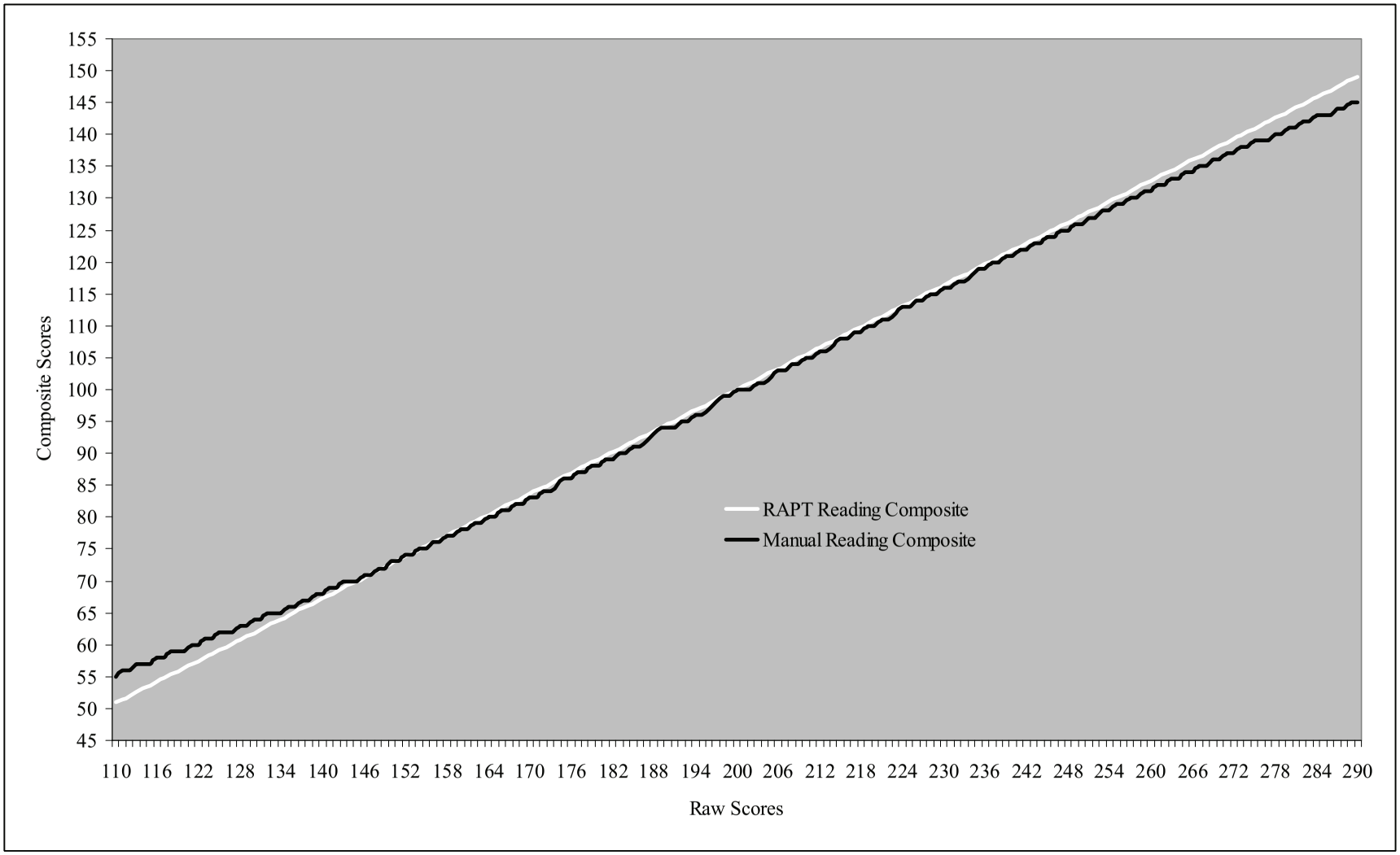
Replication of Reading Composite Standard Scores, Percentile Ranks and 90% Confidence Intervals

	RAPT RC			**WRAT-4 RC**		
		90% Confidence Interval			90% Confidence Interval	
Raw Score	Reading Composite	Low	High	Reading Composite	Low	High
.	.	.	.	.	.	.
.	.	.	.	.	.	.
192	96	91	101	95	90	100
193	96	91	101	95	90	100
194	97	92	102	96	91	101
195	97	92	103	96	91	101
196	98	93	103	97	92	102
197	98	93	104	98	93	103
198	99	94	104	99	94	104
199	99	94	105	99	94	104
200	100	95	105	100	95	105
201	101	95	106	100	95	105
202	101	96	106	100	95	105
203	102	96	107	101	96	106
204	102	97	107	101	96	106
205	103	97	108	102	97	107
206	103	98	108	103	98	108
207	104	99	109	103	98	108
.	.	.	.	.	.	.
.	.	.	.	.	.	.

As indicated in the above table, the reading composite scores calculated using RAPT methodology closely approximate those based on the WRAT-4 normative sample as provided in the professional manual. Appendix F provides the complete comparison between RC scores calculated using RAPT algorithms and those provided in the WRAT-4 testing manual, including ninety percent confidence intervals. Examination of the "look-up tables" demonstrates the capacity of RAPT methodology to replicate norms-based indices with a high degree of accuracy.

The extent of similarity is best demonstrated by examining the distribution of the calculated composite scores against those provided in the manual. This comparison is graphically displayed in Figure 7.1.

Figure 7.1 Comparison of WRAT-4 reading composite against reading composite calculated using RAPT methodologies



As indicated in this figure, the RAPT reading composite scores and those provided in the WRAT-4 professional manual were plotted against each other to visually determine the degree of deviation between indices (i.e., as demonstrated in Appendix F). The figure indicates little deviation in the middle of the distribution. RAPT and manual composites deviate slightly in the upper and lower extremes. In fact, mathematical examination of discrepancies indicted that twenty-three percent (23%) of the scores did not deviate at all, almost half (47%) deviated by only plus or minus one standard score, fourteen percent (14%) deviated by two standard scores, ten percent (10%) deviated by three standard scores, six percent (6%) deviated by four and only one percent (1%) deviated by plus or minus five standard scores. As indicated in figure 7.1, the majority of deviations occurred in the upper and lower extremes of the distribution.

In fact, this deviation in more extreme scores is an expected trend in any comparison between linear composite scores (e.g. those calculated using the proposed methodology) and scores based on a cumulative frequency distribution which has been "visually smoothed" to moderate extreme values and which is vulnerable to under-sampling in the upper and lower extremes of the normative sample (Wilkinson & Robertson, 2006). This will be discussed in greater detail below. For the purposes of the current investigation, RAPT methodology appears to replicate the WRAT-4 reading composite with some degree of accuracy particularly for less extreme scores.

7.2.2 WRAT-4 Reading Composite Reliability Coefficients

RAPT formula 5.6 was used to calculate the internal consistency coefficients of the Reading Composites for the Blue and the Green test form. A comparison of

the resulting internal consistency coefficients for the RAPT RC's and the WRAT-4 RC's are provided in Table 7.2.

Table 7. 2

Comparison of Reliability Coefficients for RAPT RC's and WRAT-4 RC's

Composite	**WRAT-4 RC Reliability Coefficient**	**RAPT RC Reliability Coefficient**
RC (Blue Form)	.96	.96
RC (Green Form)	.96	.96

As indicated in the table, the RAPT algorithms were successfully used to calculate composite reliabilities which exactly replicated those provided in the WRAT-4 professional manual.

In this demonstration, RAPT methodology was successfully used to replicate the analytical structure of a clinical composite based on an extensive normative sample. The linear composites calculated using RAPT algorithms deviated only moderately from those calculated using a normative distribution. Additionally, the reliability of the composite was successfully calculated using RAPT methodology. If existing analytical structures of widely accepted clinical and psychometric utility are replicated by the RAPT the confidence with which it may be applied to different collections of tests is increased. The current example provides just such evidence on a small scale.

7.3 Replicating WAIS-III FSIQ, PIQ, VIQ, VCI, POI, PSI and WMI

The Wechsler Intelligence Scales (i.e., WAIS; WAIS-R; WAIS-III) are by far the most widely used cognitive instruments as indicated by reviews of clinical practice (Camara, Nathan & Puente, 2000; Rabin, Barr & Burton, 2005) and as such

may be considered to constitute an acceptable standard in terms of structure, analytical and interpretive techniques, and psychometric evaluation. In the following section, clinical composite methodology will be used to replicate the reliability coefficients, intercorrelations, norms and discrepancy statistics for WAIS-III summary scores. The extent to which WAIS-III normative and ipsative statistics can be produced using the proposed algorithms is also examined.

In essence the principle here is to demonstrate that having only knowledge of the individual subtests and their psychometric properties the same high level of analysis and interpretative structure can be replicated using RAPT methodology. If this can be demonstrated convincingly then clinicians can be confident in achieving a level of analysis equivalent to that of the WAIS-III when using RAPT methodology to structure and analyse changing collections of cognitive measures.

7.3.1 Comparison of WAIS-III Manual and RAPT Summary Scores

In the following example, formulae 5.2, 5.3, 5.4, 5.6, 5.7, 5.8a, 5.9a, 5.10, 5.11a, 5.13a, 5.14 and 5.15 were used to completely replicate WAIS-III analytical structure. Linear composite scores approximating the norms based Full Scale IQ (FSIQ), Verbal IQ (VIQ), Performance IQ (PIQ), Verbal Comprehension Index (VCI), Perceptual Orientation Index (POI), Processing Speed Index (PSI) and Working Memory Index (WMI) provided in the WAIS-III administration and scoring manual (Wechsler, 1997b) were calculated and RAPT methodology was used to approximate the composite scores, percentiles and confidence intervals provided in the same manual for interpretation of these constructs, to calculate the internal consistency of the resulting composite scores and to provide analysis of the significance and abnormality of discrepancies between composite scores provided by

WAIS-III analytical methods. Subtest combination replicated that indicated by factor analytic research provided in the technical manual and the subtests included in RAPT composites were those indicated by WAIS-III analytical structure and all normative information was obtained from the technical manual (Wechsler, 1997a). The results of calculations are provided as look-up tables in appendix G which replicate those provided in the WAIS-III administration and scoring manual (Wechsler, 1997b). Examination of these tables indicates that RAPT methodology can be used to accurately replicate WAIS-III composite scores.

A section of results from these analyses for the processing speed index (PSI) is compared with the normative PSI provided in the administration and scoring manual tables in Table 7.3 (Wechsler, 1997b). As indicated in the table, the RAPT PS composites associated with sums of subtest scaled scores closely approximate the PSI provided in the technical manual. Lookup tables for FSIQ, VIQ, PIQ, VCI, POI and WMI and the full table for PSI are provided in appendix G indicating a similar capacity of RAPT methodology to replicate WAIS-III summary scores, when these results are compared with look-up tables provided in the administration and scoring manual (Wechsler, 1997b).

Table 7.3

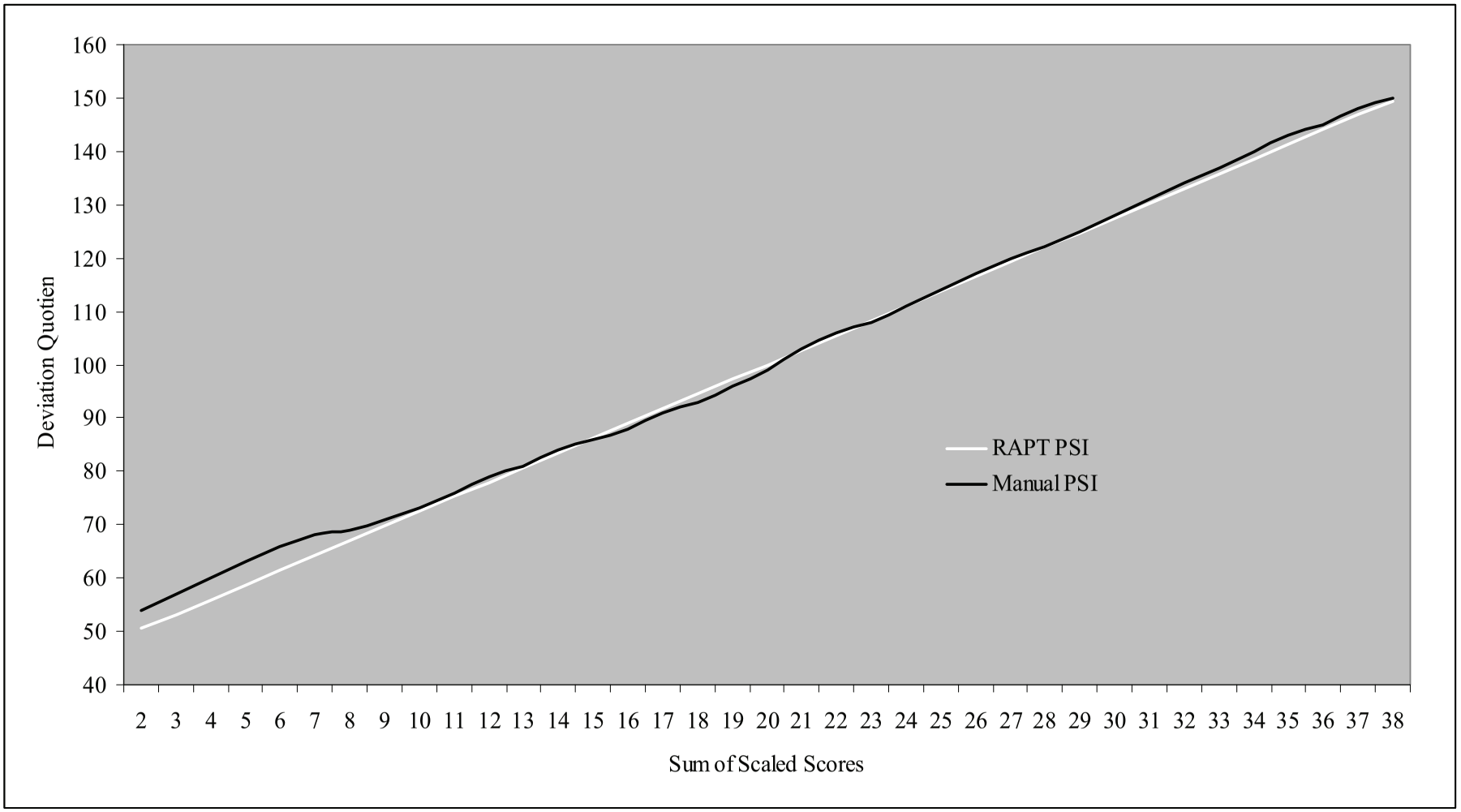
WAIS-III PSI and RAPT PSI Composites and 90% Confidence Intervals

PSI			90% Confidence Interval	
SS	WAIS-III	RAPT	WAIS-III	RAPT
.	.	.	.	.
10	73	72	69-84	68-84
11	76	75	71-86	70-86
12	79	78	74-89	73-89
13	81	81	76-91	75-91
14	84	83	78-93	77-93
15	86	86	80-95	80-96
16	88	89	82-87	82-89
17	91	92	85-100	85-101
18	93	94	86-101	87-103
19	96	97	89-104	90-106
20	99	100	92-107	92-108
21	103	103	95-110	94-110
22	106	106	98-113	97-113
23	108	108	99-115	99-115
24	111	111	102-117	102-118
25	114	114	105-120	104-120
26	117	117	107-123	107-123
27	120	119	110-125	109-125
28	122	122	112-127	111-127
29	125	125	114-130	114-130
30	128	128	117-132	116-132
31	131	130	120-135	119-135
32	134	133	122-137	121-137
33	137	136	125-140	120-140
34	140	139	128-143	126-142
35	143	141	130-145	128-144
.	.	.	.	.

Calculated summary scores were then compared graphically with norms-based summary scores to evaluate the degree of similarity between linear (RAPT) and normative (WAIS-III) indices. The distributions of RAPT composites and WAIS-III composites are plotted to graphically compare the extent to which RAPT algorithms successfully replicate WAIS-III index and summary scores. Figure 7.2 displays a graphical comparison of the WAIS-III PSI and RAPT PSI, plotted according to the sum of scaled scores. Examination of this graph demonstrates that the RAPT PSI (plotted in white) closely approximates the PSI composite provided in the WAIS-III manual (plotted in black). In fact, very little deviation occurs, except in the lower end of the distribution.

As with the WRAT-4 reading composite, it is notable that the linear transformation employed in RAPT algorithms results in a much smoother distribution of scores which adheres to the expectations of a normal distribution. As with the WRAT-4, the normal distribution was approximated in the WAIS-III standardisation by "visual smoothing" and adjustment of the sampling distribution to approximate linear transformation where sampling in the extreme ends of the distribution was constrained by under-sampling (Wechsler, 1997a).

Figure 7.2 Comparison of WAIS-III PSI and RAPT PSI



This comparison, again, highlights a salient issue in psychometric assessment concerning the veracity of various normative methods which, at this stage, warrants some discussion. WAIS-III summary scores are based on uniform norming, in which cumulative frequency distributions of actual sums of scaled scores in the standardisation sample are used to produce scale and index scores (Wechsler, 1997a). Such methodology is vulnerable to the small sample sizes in the extremes of the distribution, which provide little actual data upon which to base normative frequencies. In contrast, RAPT algorithms base summary scores on linear transformation which applies normative means and standard deviations to the development of a smooth distribution of scores closely approximating those expected from a normally distributed population with the same parameters of deviation and central tendency.

While typical practice treats the normative scores presented in test manuals as somewhat sacrosanct, in fact various mathematical techniques are likely to be employed to modify uniform distributions with the aim of ensuring that the sample distribution more closely approximate the normal distribution and avoids the possible confounds associated with sampling. Where adequate sampling is unachievable, such as is likely at extreme test scores, test developers may employ various modifications to normative frequencies to increase the degree to which the sample accurately approximates the normal distribution and facilitates the development of frequency based scores. For the WAIS-III for example, the norming process upon which summary scores are based included “smoothing and normalising” of the cumulative frequency distribution and successive adjustments “based on computerized smoothing and visual inspection of the distributions” (Wechsler, 1997a, p. 42), a somewhat ambiguous description of methodology. In fact,

approximation of the normal distribution may arguably be more exactly achieved using linear norms based only on measures of normative central tendency and average deviation.

Closer examination of this issue is unnecessary at this stage. It is safe to say that definitive conclusions regarding which distributions provide a "better" approximation of reality are still unclear, though perhaps linear transformation adheres most closely to the normative "ideal" and clearly avoids any vulnerability to errors of sampling or "clinical" judgements required to manually modify the distribution of standardisation data. For the purposes of the current investigation, it is likely that discrepancies between RAPT composites and those provided in the Wechsler manual deviate largely due to this methodological difference. Such deviation provides little evidence against the linear methodology employed by RAPT. It can be concluded that despite differences in normative technique, RAPT algorithms may be used to calculate a PSI which approximates the normal distribution with at least the accuracy achieved through use of the WAIS-III uniform norms.

The graphs that follow (Figure 7.3 – 7.8) show the remaining linear summary scores (i.e., FSIQ, VIQ, PIQ, VCI, POI and WMI) plotted against those provided in the administration and scoring manual. These figures conclusively indicate that while deviation from linear transformation occurs in the manual summary scores above index scores of approximately 130 and below those of approximately 60, and while the manual norms are artificially constrained in the upper and lower extremes, the summary scores calculated using RAPT algorithms closely approximate those found in the WAIS-III administration and scoring manual (Wechsler, 1997b).

Figure 7.3 Comparison of RAPT FSIQ and WAIS-III Manual FSIQ

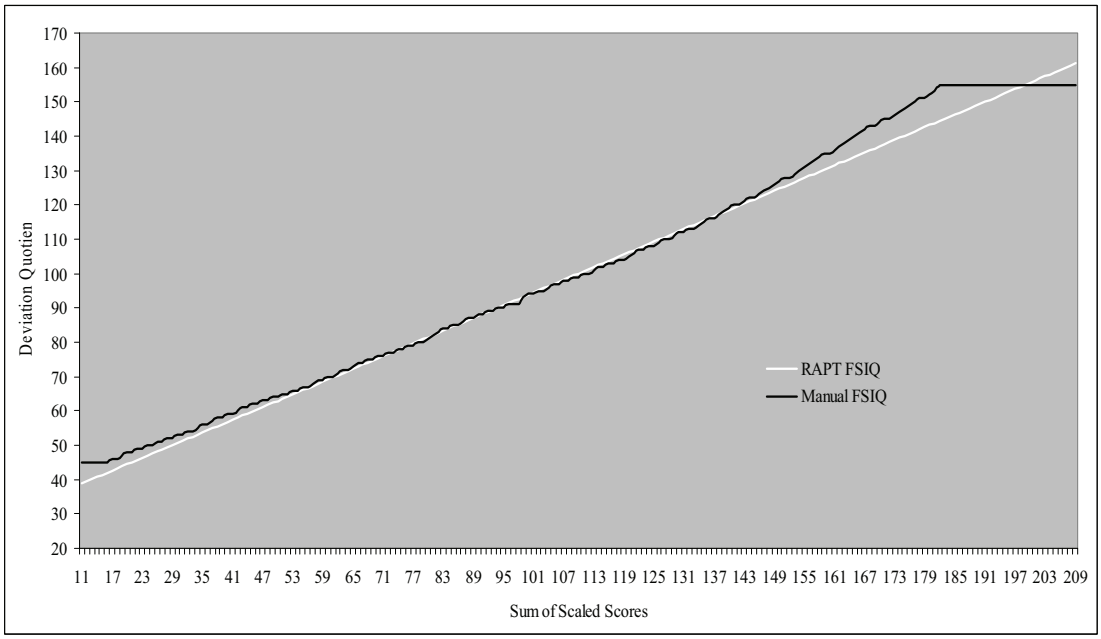


Figure 7.4 Comparison of RAPT VIQ and WAIS-III Manual VIQ

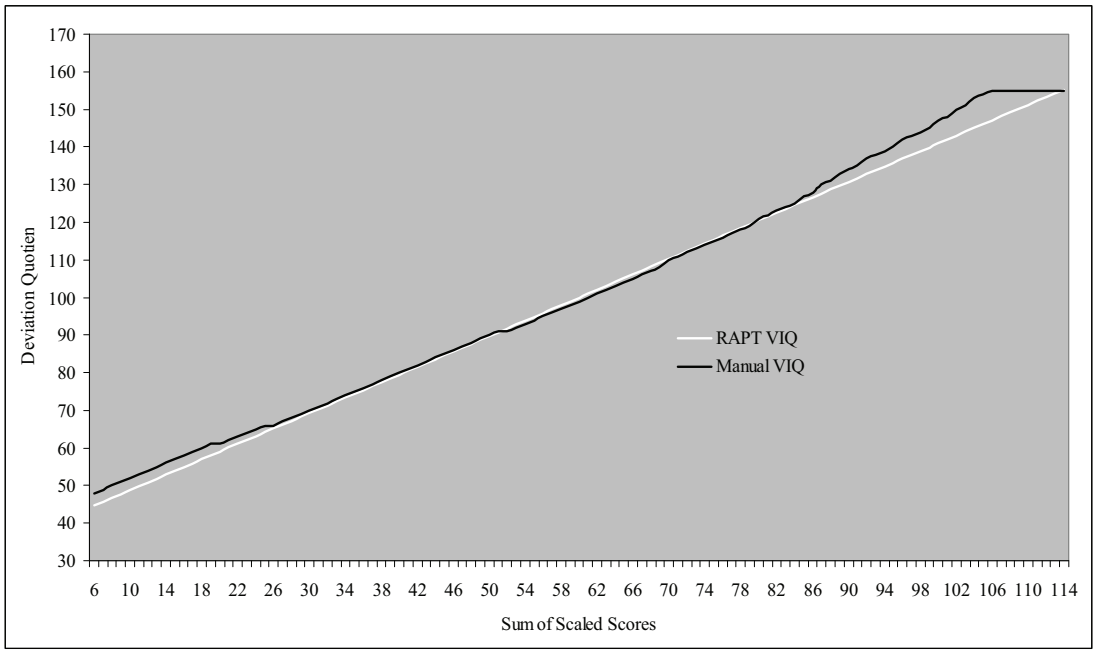


Figure 7.5 Comparison RAPT PIQ and WAIS-III Manual PIQ

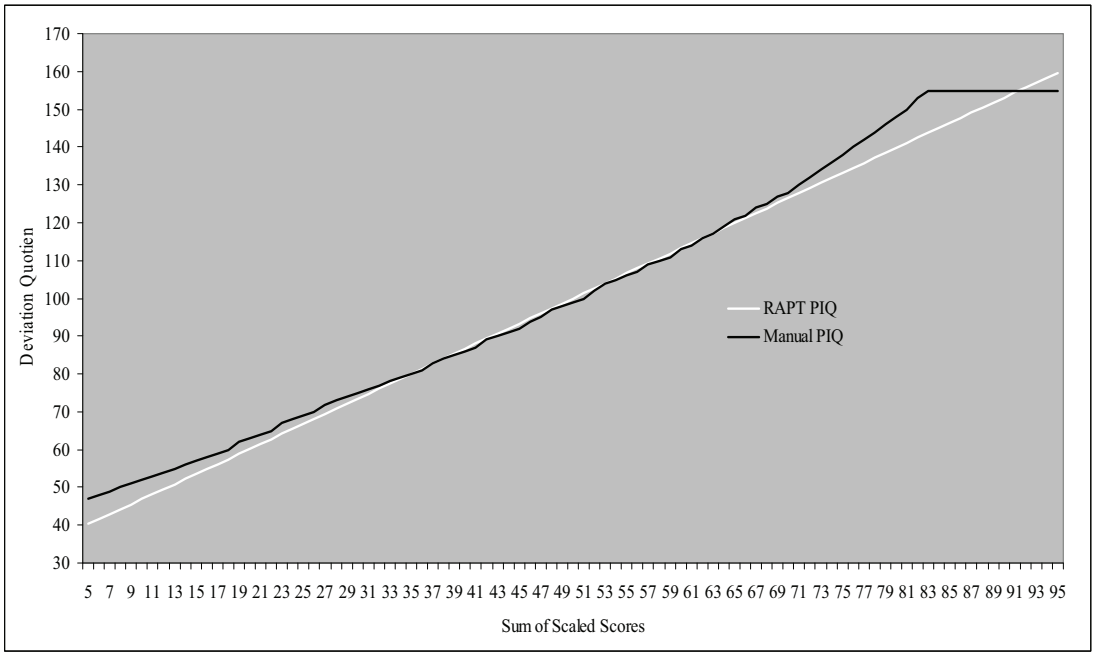


Figure 7.6 Comparison of RAPT VCI and WAIS-III Manual VCI

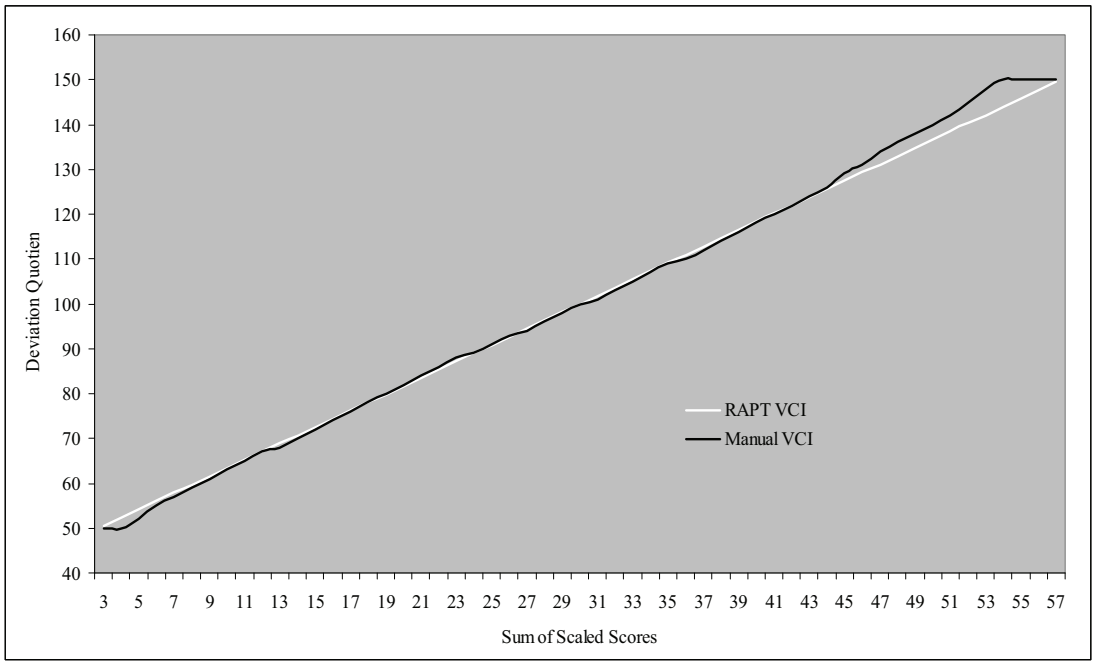


Figure 7.7 Comparison of RAPT POI and WAIS-III Manual POI

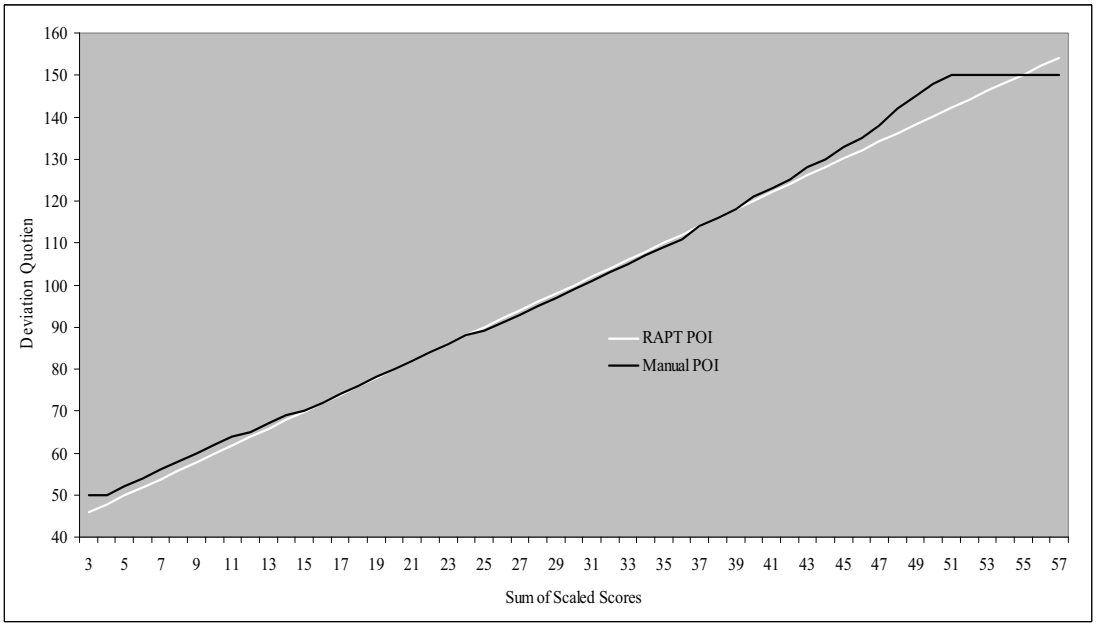
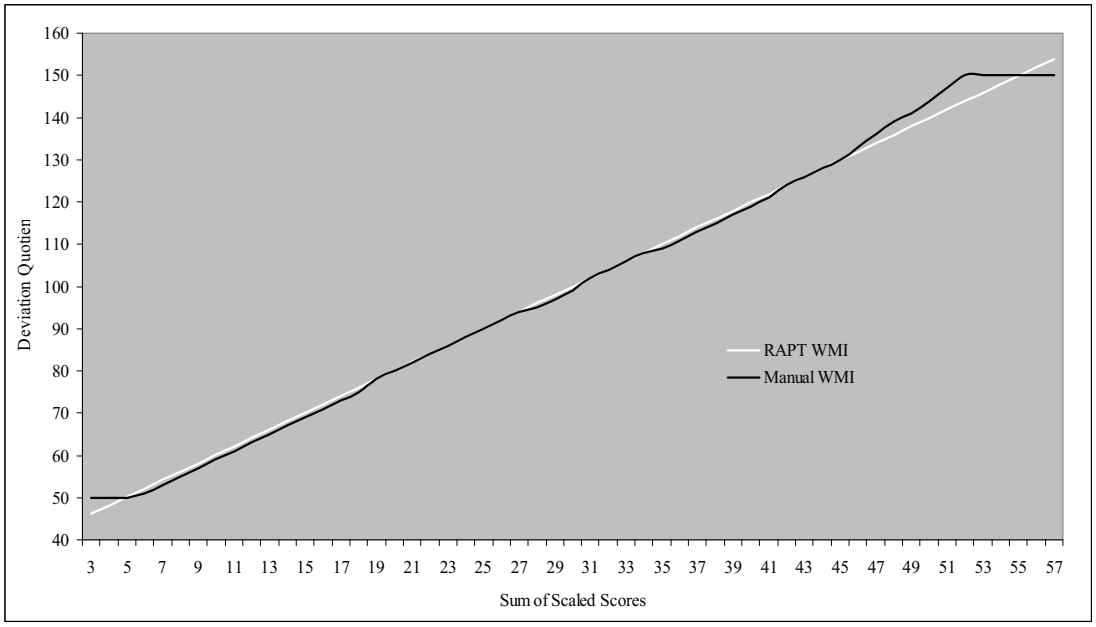


Figure 7.8 Comparison of RAPT WMI and WAIS-III Manual WMI



To further investigate the relationship between RAPT and WAIS-III summary scores, differences in these composites were compared numerically and presented in Table 7.4. This table indicates the percentage of deviations found between RAPT and WAIS-III summary scores.

Table 7.4

Percentage of Deviation between WAIS-III Summary Scores and Those Approximated using RAPT Methodology

	FSIQ	**VIQ**	**PIQ**	**VCI**	**POI**	**WMI**	**PSI**
No Difference	23%	30%	16%	44%	25%	35%	32%
Within 1 point	51%	49%	45%	75%	58%	78%	78%
Within 2 points	67%	59%	57%	84%	80%	85%	84%
Within 3 points	75%	79%	71%	91%	84%	89%	84%
Within 4 points	81%	85%	76%	95%	91%	96%	97%
Within 5 points	85%	90%	84%	96%	91%	98%	100%
Within 6 points	89%	93%	89%	100%	95%	100%	
Within 7 points	94%	97%	93%		96%		
Within 8 points	96%	100%	95%		100%		
Within 9 points	99%		97%				
Within 10 points	99%		98%				
Within 11 points	100%		100%				

Overall, analyses of discrepancies indicated that RAPT methodology approximated WAIS-III index scores with minimal deviations. Specifically, RAPT composites deviated a maximum of ± 11 points from those provided in the manual with a large majority of comparisons showing only relatively small discrepancies (i.e., ± 4 points) between the indices associated with various sums of scaled scores. Approximation of FSIQ and PIQ scores were the least accurate with scores deviating

a maximum of 11 points above or below. In contrast approximation of several index scales were very accurate. For example, all deviations between the RAPT PSI and WAIS-III PSI were within five points and the majority (i.e., more than 80%) were within two points for this index. VCI and WMI indices deviated a maximum of 6 points, while VIQ and POI deviated a maximum of 8. As discussed, such deviations express the extreme discrepancies produced by normative techniques (i.e., linear or uniform) in the upper and lower ends of the distribution and the discrepancies associated with more moderate scores were minimal, providing evidence of the robustness of RAPT methodology to provide interpretable summary scores.

7.3.2 Comparison of WAIS-III Manual and RAPT Reliability Coefficients

RAPT formula 5.6 was used to calculate the reliability coefficients for WAIS-III summary and index scores (Table 7.5).

Table 7.5

Replication of WAIS-III Composite Reliability Coefficients using RAPT Methodology

	Calculated Reliability	**Technical Manual Reliability**	**Difference**
FSIQ	.98	.98	0
VIQ	.97	.97	0
PIQ	.94	.94	0
VCI	.96	.96	0
POI	.93	.93	0
WMI	.94	.94	0
PSI	.88	.88	0

As indicated in the table, the internal consistency coefficients of WAIS-III summary scores were perfectly replicated using the normative algorithms. While this would appear to be a resounding success for the RAPT methodology it is, in fact, unsurprising given that the formulae used in the RAPT approach are identical to those employed calculating the reliability coefficients presented in the WAIS-III technical manual.

It is important to recognise, however, that the use of the RAPT algebraic formulae do not reflect deviations from standard practice but are employed routinely in test construction and development and are merely being employed in RAPT to provide the same level of analysis to the selection and analysis of semi-flexible batteries. These results indicate that clinicians could use RAPT methods to calculate composite reliabilities with at least the degree of confidence warranted by use of the WAIS-III technical manual.

7.3.3 Comparison of WAIS-III Manual and RAPT Intercorrelations between Summary and Index Scores

To determine if intercorrelations between WAIS-III summary scores could be successfully replicated comparison was made with those calculated using formula 5.7. The resulting intercorrelation matrices are provided in Table 7.6. As indicated in the table, the intercorrelations between WAIS-III technical manual summary scores which provide the "average correlation across all 13 age groups computed with Fisher's z transformation" (Wechsler, 1997a) are well replicated using RAPT algorithms. Intercorrelations between summary scores deviated a maximum of .03 and the majority were accurately replicated using the proposed methodology.

Table 7.6

Replication of WAIS-III Intercorrelations using RAPT Methodology

	RAPT Intercorrelations				
	PIQ	*VCI*	*POI*	*WMI*	*PSI*
VIQ	.75	.95	.71	.82	.58
PIQ		.71	***.95***	***.67***	***.76***
VCI			.67	.65	.53
POI				.63	.60
WMI					***.58***

	WAIS-III Manual Intercorrelations				
	PIQ	*VCI*	*POI*	*WMI*	*PSI*
VIQ	.75	.95	.71	.82	.58
PIQ		.71	***.94***	***.65***	***.75***
VCI			.67	.65	.53
POI				.63	.60
WMI					***.55***

As in the WAIS-III and WMS-III technical manual, some calculated correlations are inflated due to shared subtests. The RAPT algorithms successfully replicated those provided in the WAIS-III manual in all but four instances (i.e., PIQ-POI, PIQ-WMI, PIQ-PSI and WMI-PSI), however, these deviations were minimal at .01, .02, .01 and .03 respectively. These results suggest that RAPT methodology may be used with a justifiable degree of confidence to investigate relationships between the composite measures of a cognitive battery.

7.3.4 Comparison of WAIS-III Manual and RAPT Summary Score Discrepancies

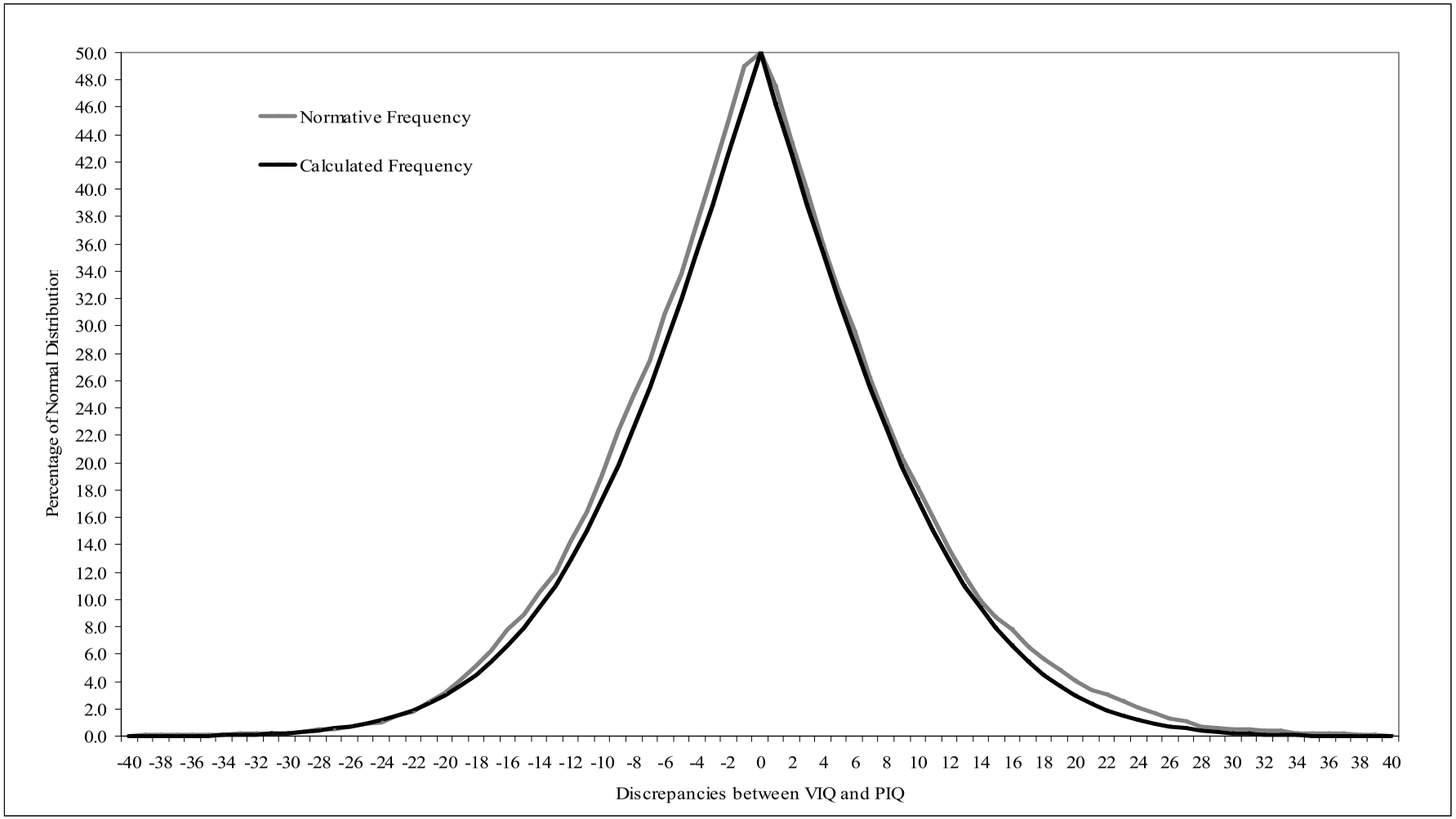
To determine if inappropriate error would be introduced by using RAPT algorithms to calculate the abnormality of a discrepancy, the normative frequency associated with discrepancies obtained using algorithms (formula 5.15) was compared with the actual frequencies associated with discrepancies in the WAIS-III standardisation sample as provided by Tulsky, Rolfhus and Zhu, (2000) and the cumulative frequencies calculated according to a heuristic proposed by Tulsky, Zhu and Vasquez (1998) and Sattler and Ryan (1998) in which the manual frequencies

are divided by two. Results indicated that RAPT methodology produced highly similar results to those proposed in previous research using the WAIS-III standardisation sample.

Results for the comparison between VIQ and PIQ, for discrepancies from -40 to +40, are graphically displayed in Figures 7.9 below which demonstrates a comparison of the normative frequencies obtained from the WAIS-III standardisation sample derived by Tulsky, Rolfhus and Zhu (2000) and those obtained using the RAPT discrepancy analysis algorithm.

The discrepancy frequencies between VIQ and PIQ calculated using RAPT methodology closely approximated the actual frequencies associated with discrepancies in the WAIS-III standardisation sample (Tulsky, Rolfhus & Zhu, 2000). In fact, the frequencies of discrepancies between VIQ and PIQ calculated using RAPT algorithms deviated a maximum of ± 2.76 from the actual frequencies in the normative sample and a maximum of ±2.06 from the frequencies obtained by dividing manual frequencies by two.

Figure 7.9 Comparison of the frequencies of VIQ-PIQ discrepancies in the WAIS-III normative sample with those calculated using RAPT



Similarly positive results were obtained for the discrepancies between other index scores, as indicated in Table 7.7.

Table 7.7

Deviation between Frequencies Associated with RAPT Discrepancies and Normative Discrepancies

	Maximum deviation between RAPT and Normative Sample
VCI – POI	± 3.02
VCI – WMI	± 5.38
VCI – PSI	± 2.12
POI – WMI	± 2.06
POI – PSI	± 2.63
WMI – PSI	± 3.17

These comparisons are graphically displayed in Figures 7.10 – 7.15 below which again plot the normative frequencies obtained from the WAIS-III standardisation sample (Tulsky, Rolfhus & Zhu, 2000; n = 2450, or weighted n = 1250 for comparisons with WM) and those obtained using formula 5.15 of the RAPT methodology.

Figure 7.10 Comparison of the frequencies of VCI-WMI discrepancies in the WAIS-III normative sample with those calculated using RAPT

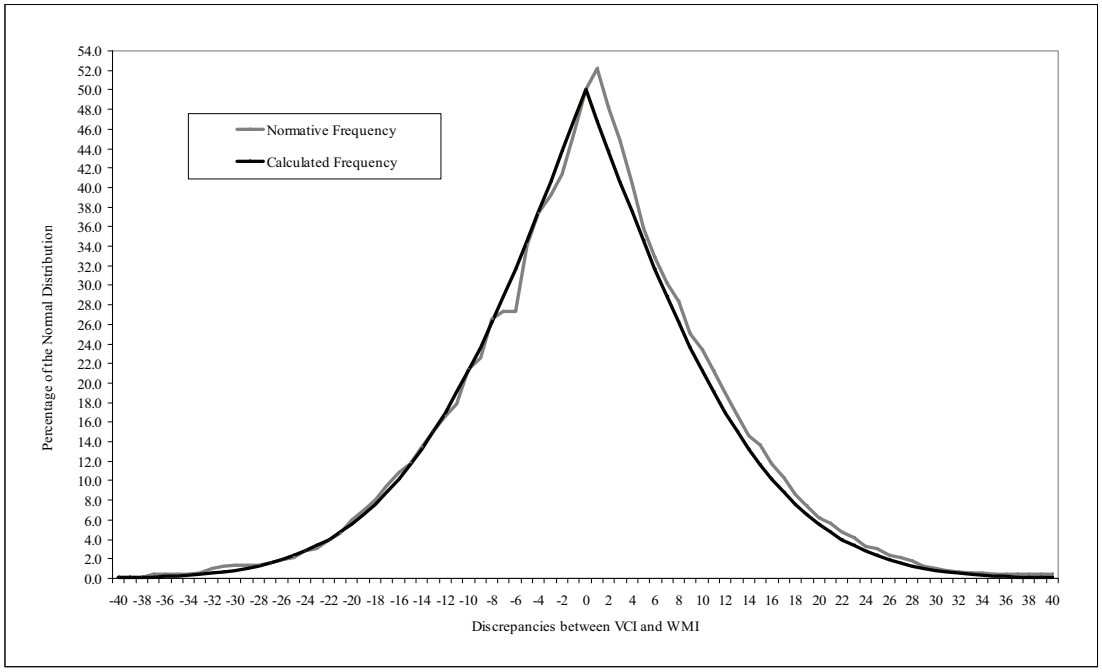


Figure 7.11 Comparison of the frequencies of VCI-POI discrepancies in the WAIS-III normative sample with those calculated using RAPT

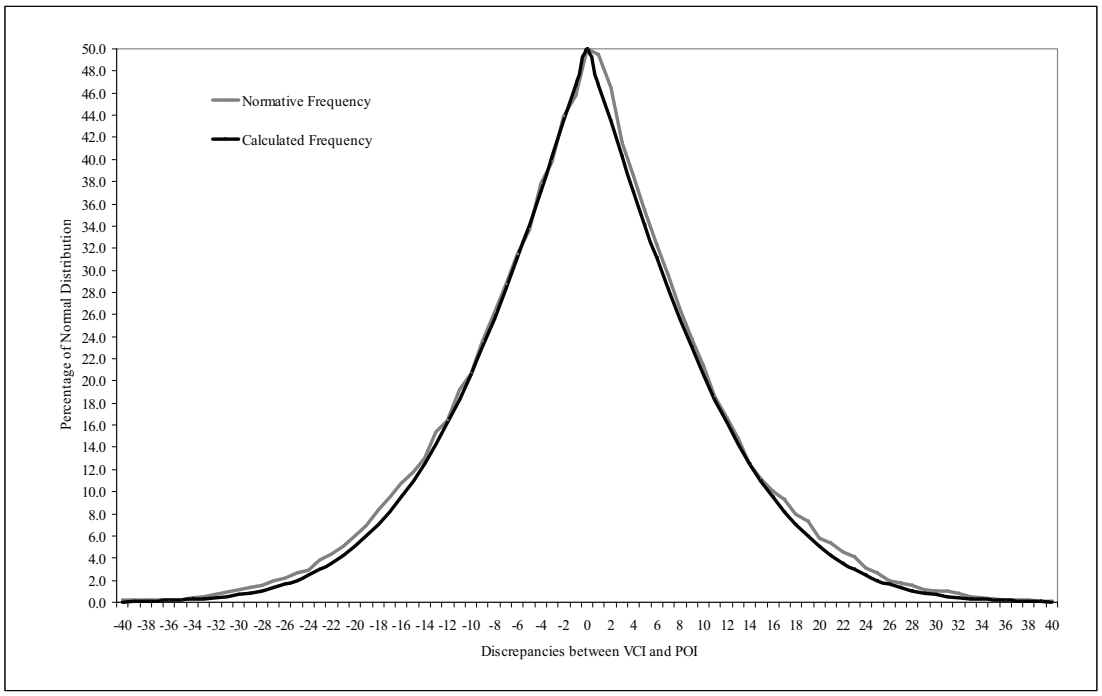


Figure 7.12 Comparison of the frequencies of VCI-PSI discrepancies in the WAIS-III normative sample with those calculated using RAPT

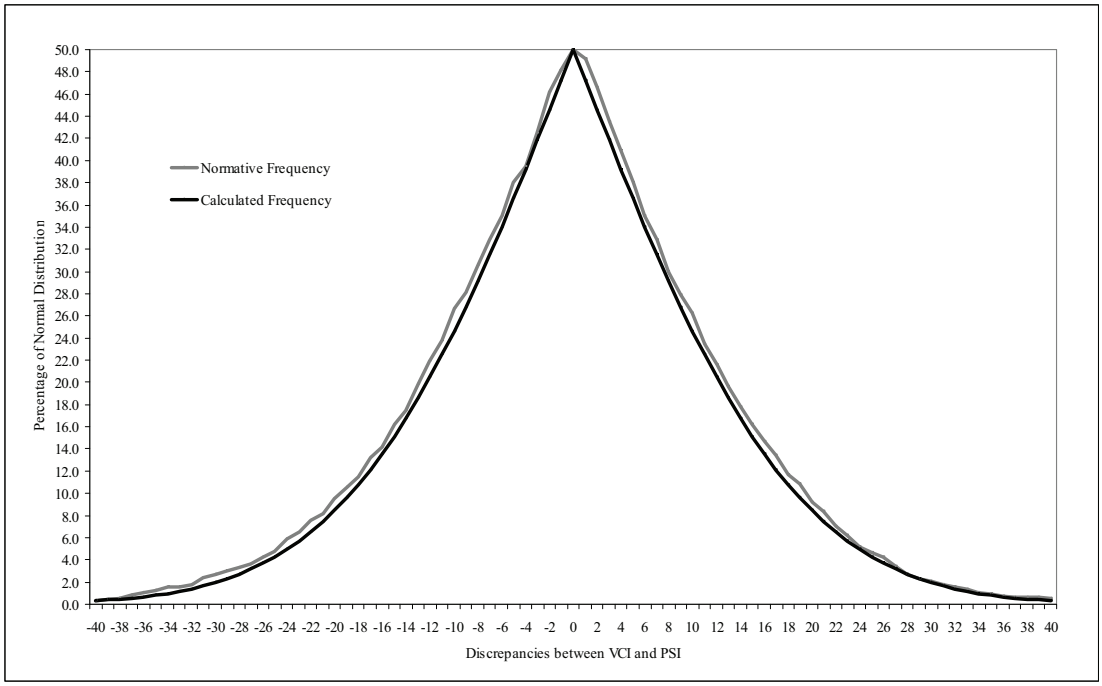


Figure 7.13 Comparison of the frequencies of POI-WMI discrepancies in the WAIS-III normative sample with those calculated using RAPT

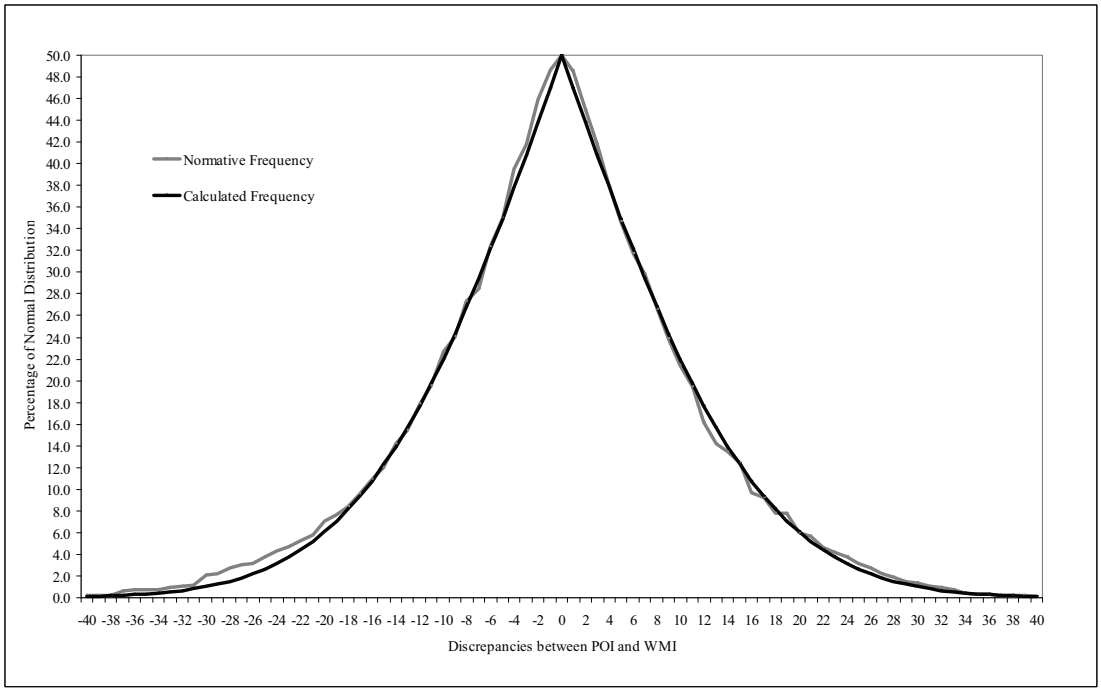


Figure 7.14 Comparison of the frequencies of POI-PSI discrepancies in the WAIS-III normative sample with those calculated using RAPT

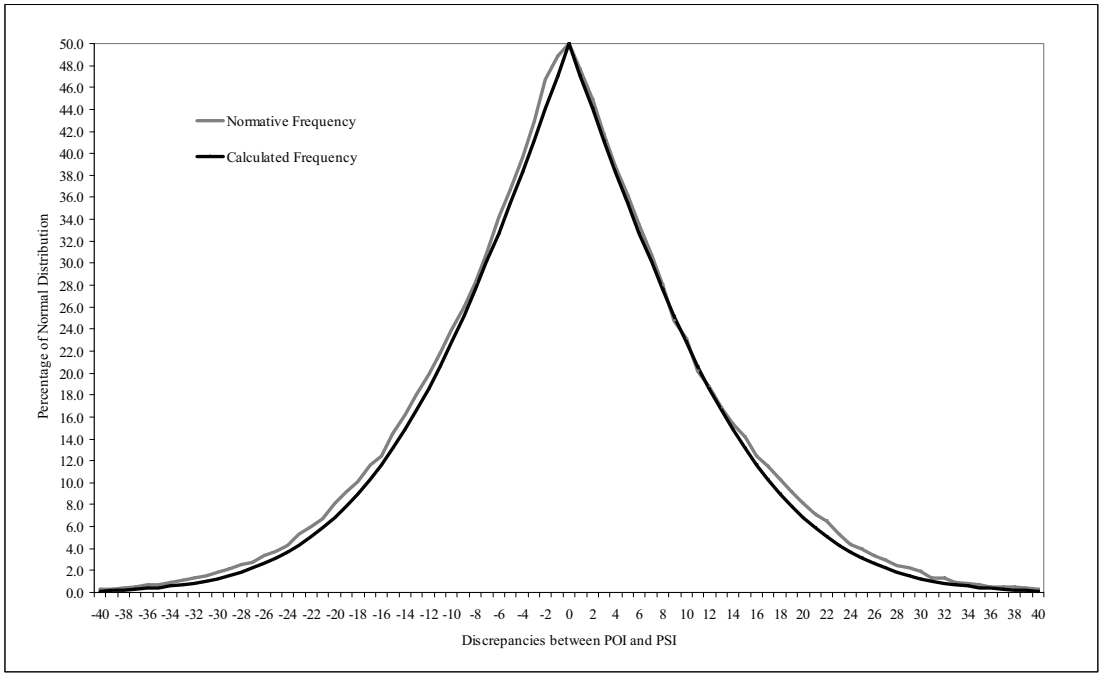
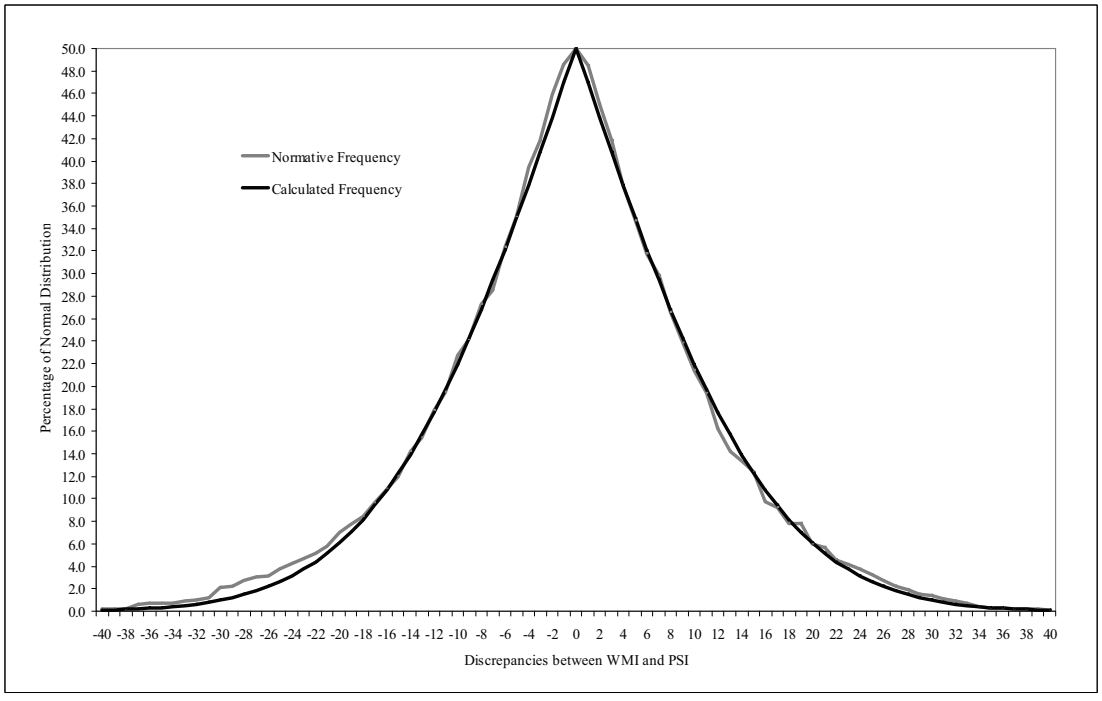


Figure 7.15 Comparison of the frequencies of WMI-PSI discrepancies in the WAIS-III normative sample with those calculated using RAPT



These figures indicate a close approximation of normative frequencies using RAPT methodology, providing confidence in use of this algorithm to facilitate analysis of profile discrepancies within a battery of composite scores.

Use of collections of tests, such as the WAIS-III or WRAT-4, are commonly accepted by clinicians as constituting sound psychometric and clinical practice, as well as reducing the complexity of data obtained from a typical cognitive battery. The analytical and interpretative framework of such "batteries" facilitates sound clinical judgements with due consideration of psychometric factors and provides some resolution of the complexity typically attending semi-flexible collections of tests. The previous demonstrations indicated the capacity of RAPT methodology to successfully replicate the analytical and interpretive structures of the WRAT-4 and WAIS-III summary scores with at least the accuracy available in technical manuals. The capacity of RAPT methodology to accurately re-create analyses possible using the existing analytical framework of these tests should allow the clinician to employ RAPT methods to a semi-flexible collection of measures with a justifiable degree of confidence in the likely accuracy of resulting clinical inferences. It is this capacity which constitutes the unique contribution of the RAPT method.

7.4 A Word Regarding the Rohling Interpretive Method

It would be remiss to suggest that these issues are not of interest to or being undertaken by other researchers. At the forefront of these endeavours is Miller and Rohling's explication of the Rohling Interpretive Method (RIM; Miller & Rohling, 2001; Rohling, Langhinrichsen-Rohling, & Miller, 2003). The RAPT and RIM share highly similar objectives and both aim to improve the clinical usage of flexible and semi-flexible batteries by the application of actuarial method. Specifically, authors

of the RIM aim to minimise human judgement errors endemic to flexible battery usage by providing summary results "analogous to those generated in a fixed-battery approach" and facilitating quantitatively-based comparisons at the overall, domain and test levels, within a comprehensive clinical battery (Rohling & Miller, 2004). The RIM is designed to apply to any representative changing collection of measures and to be useful in the assessment of any individual client and as such improves the defensibility of battery results in light of criticisms to the professions discussed in previous sections. The RIM has been demonstrated to improve diagnostic accuracy and to facilitate a systematic approach to treatment planning (Rohling, Miller & Langhinrichsen-Rohling, 2004; Rohling, Williamson, Miller & Adams, 2003).

It would seem not only appropriate but warranted to compare the RAPT methodology with the RIM at this stage. However, the theory, steps and application of the RIM differ fundamentally from those of the RAPT both in structure and analytical ethos. For example, while both methodologies advocate the calculation of composite measures, these summary scores differ in terms of composition, psychometric characteristics and intended role in subsequent clinical inferences. Similarly, while the RIM suggests a system of profile analysis in which the significance of discrepancies between summary scores are thoroughly investigated, RAPT suggests that clinicians choose between normative and ipsative analytical methods and focus on developing a thorough understanding of the reliability associated with test combinations.

While both methods propose a means of organising the complex data resulting from administration of a semi-flexible battery, they employ widely differing actuarial methods which render direct comparison not only unwieldy but

inappropriate. The current section can simply acknowledge the contribution of RIM objectives to the development of the current RAPT methodology.

7.5 Conclusions

The current chapter has provided empirical evidence of the applicability of RAPT methodology to the clinical use of cognitive batteries. The RAPT method has been shown to capitalise on the psychometric and practical strengths of composite methodology and to replicate well-known and widely employed analytical structures.

Composite scores provide a means of simplifying interpretation and formalising validated structure and are frequently employed as summary scores, cut-off scores or diagnostic indicators within standardised test collections. Basing test selection, analysis and inference on a system of composite scores based on empirically validated cognitive domains provides a formalised means of balancing clinical and psychometric factors when a semi-flexible battery is used and can be readily applied to improving the psychometric quality of clinical practices. Despite their advantages and the availability of the necessary methodology, composite scores are rarely employed at the level of the semi-flexible cognitive battery. The current chapter, however, has demonstrated the capacity of RAPT methodology to provide just such a structure even when the battery remains flexible.

In the current chapter, RAPT methods have been used to successfully replicate the analytical structure and indices used in WRAT-4 and WAIS-III batteries. Fixed measures such as WRAT-4 and WAIS-III are widely considered to exemplify psychometrically sound practices, in part due to their efficient application of composite methodology and empirically validated structure. Knowing only the mean, standard deviation and internal consistency coefficient for each WRAT-4 or

WAIS-III subtest as well as the intercorrelations between tests in the battery, summary scores, composite reliabilities, composite intercorrelations and discrepancy analyses can be accurately reproduced using RAPT. Given clinicians' ubiquitous reliance on Wechsler scales, the capacity of RAPT methodology to accurately replicate WAIS-III summary scores should facilitate confidence. Specifically, clinical conclusions based on measures beyond the somewhat limited scope of WAIS-III subtests may be analysed using RAPT methods providing improvements in terms of clinical flexibility and psychometric robustness.

The RAPT methodology capitalises on the psychometric strengths of composite scores and has a demonstrable ability to replicate well-known analytical structures. The current chapter provided evidence that clinicians may use the suggested composite methodology to create composites of any empirically valid combinations, to evaluate the associated reliability, and to conduct normative and ipsative analyses with at least the confidence warranted by use of existing fixed-battery composite methodology. It is hoped that this explication provides clinicians with sufficient empirical evidence of the value of RAPT methodology to improve the ways in which changing batteries are used clinically.

CHAPTER EIGHT

USING RAPT METHODOLOGY

8.1. Introduction

RAPT methodology provides a sound method of integrating psychometric theory with the practical constraints of clinical practice. In doing so, RAPT can be used to provide answers to several challenges to semi-flexible battery usage raised in the literature. Previous chapters have demonstrated the capacity of RAPT to improve the stability, accuracy and meaningfulness of behavioural quantification using semi-flexible batteries of cognitive tests. These demonstrations provide evidence regarding the capacity of the proposed methodology to replicate widely used and accepted measures of cognitive functioning and indicate that the proposed methodology constitutes an improvement, in terms of efficiency, reliability, validity and adaptability, to current methods of quantifying cognitive behaviours. Throughout chapters five, six and seven the focus has been on the theoretical capacity of RAPT methodology. A practical demonstration of how RAPT may be flexibly applied to various tasks undertaken during psychometric testing is now warranted.

Perusal of chapter five highlights the complexity of using RAPT methodology when calculations are accomplished by hand. Given the ever present time constraints of clinical practice clinicians may find extensive application of the methodology prohibitive. However, as with many actuarial methods, RAPT algorithms may be readily entered into spreadsheet software to facilitate efficient usage. This chapter will demonstrate the ease with which RAPT methodology may be automated using an electronic spreadsheet, such as Microsoft Office Excel 2003. Examples of spreadsheets are included with this thesis to demonstrate the simplicity

of using RAPT methodology when formulae are pre-programmed. Examples in this chapter will make use of these programmed spreadsheets to demonstrate the ease with which RAPT may be applied to answering various clinical questions when analyses are computerised.

RAPT methodology may be used in clinical practice at varying levels, ranging from the substitution of alternate subtests into existing composite scores to investigation of the impact of new research on a customary battery. In addition, the methodology may be used by teachers of psychometric testing theory as a means of demonstrating psychometric principles. At the most extensive level, the RAPT may be applied to the structure, development and interpretation of a battery specifically suited to a particular clinical usage and based on factor analysis research. The various levels at which RAPT may be applied are summarised in Table 8.1. As indicated in the table the levels at which RAPT may be applied are “cumulative” in nature: each level incorporates all components of the previous levels to achieve an increasingly complete degree of application.

Table 8.1

Application of Composite Methodology to Clinical Practice

Level	Type of Application
Level Zero *"Just looking thanks. . ."*	Used to evaluate the consequences of test substitution or of alternate cognitive structures in the battery. At this level the clinician is not interested in generating numeric composites but wishes to understand whether test combinations are increasing or decreasing the potential for error.
Level One *"I've just got this little problem. . ."*	Brief use of specific formulae to answer a specific clinical question or to slightly modify an existing test framework.
Level Two *"I'll have one of those, and one of these, and . . ."*	Application of formulae to provide alternate composites in addition to those already used using familiar methodology.
Level Three *"In for a penny, in for a pound!"*	Use of methodology at all levels of battery construction, analysis and interpretation

As indicated in the table, RAPT methodology may be used, with little impact on actual clinical analyses, to evaluate the psychometric consequences of various test combinations (Level Zero). For example, as demonstrated in previous chapters reliability algorithms may be used to evaluate the psychometric stability associated with various test combinations. At this level, the aim is not to modify an existing battery but to better understand the psychometric consequences associated with test combinations. The fundamentally investigative nature of such application is implicit in subsequent levels, however, clinicians, researchers or teachers may choose to use RAPT algorithms solely to better understand the psychometric characteristics of their batteries rather than to modify existing test combinations.

Of course, evaluation of test combinations readily lends itself to modification of existing test groups to better capitalise on potential psychometric or clinical strengths. To this end, RAPT methodology may be used sparingly to add or subtract subtests from known analytical structures at a Level One application. For example, composites within the WAIS-III, WISC-IV, WMS-III or other similarly analysed cognitive batteries may require modification to accommodate unavoidably non-standardised administrations or to substitute a more reliable or valid subtest for a specifically compromised measure. The advantage of this level of application is that it suits the practice of a majority of clinicians who may need to modify test selection for various practical reasons, without substantial changes to typical battery usage.

RAPT may also be used, at the next level of application (Level Two) to modify the composition of composites based on validity research to a more substantial extent. At this level, RAPT is not simply used to add or subtract subtests from familiar composites, but to structure and analyse new combinations of test scores. RAPT may be used to calculate composite scores for which no existing

analytical framework exists, but which are suggested by validity research, theories of cognitive functioning, increased reliability or clinical factors.

Finally, as has been discussed in preceding chapters, RAPT methodology is well suited to guide the formulation and use of an entire cognitive battery. At the most extreme application, (Level Three) the test user would formulate a structure of cognitive functioning specifically tailored to suit a particular testing environment. Such a structure would form the basis for test combinations which may then be interpreted according to RAPT guidelines. This would allow clinicians ultimate flexibility in addressing clinical concerns, while still attending to the need for psychometric robustness. At this level of application, RAPT becomes more similar to other battery methods, such as the Wechsler scales, which use comparable analytical formulae. However, as discussed in previous chapters, the RAPT method differs markedly from such fixed batteries because of its fundamentally flexible nature. At the third level of application, RAPT may be used to structure an entire flexible battery which may be subsequently modified when necessary.

While the preceding section has reviewed the scope of RAPT application, it is best elucidated through practical examples of application at each of the discussed levels. The following sections provide examples of analyses at each of these levels with the aim of demonstrating the use of RAPT to successfully address clinically based questions or dilemmas. Specifically, the following examples demonstrate the use of RAPT methodology to:

a) Add and subtract individual subtests from existing clinical composites;
b) Restructure an area of cognitive functioning to form a new clinical composite and compare these with existing composites; and,

c) Structure an entire battery specifically suited to a clinical practice and potential client population.

All calculations will make use of the composite calculator spreadsheets provided with the thesis to illustrate the ease with which these investigations may be conducted using RAPT methodology.

8.2 Substituting Subtests into Existing Composites

RAPT may be used to remove a specifically confounded subtest from an existing analytical structure or to increase the efficiency of a composite structure. For example, composite methodology may be used to replace a confounding subtest within the WAIS-III Working Memory Index (WMI). WMI consists of the subtests Arithmetic (AR), Digit Span (DSp) and Letter-Number Sequencing (LNS) and is intended as a measure of the test takers capacity to attend and concentrate on cognitive tasks. In the presence of a specific dyscalculia, however, the WMI may be substantially reduced due to poor performance on the Arithmetic subtest which is not representative of the individual's overall ability to attend and concentrate. Additionally, clinicians frequently administer both the WAIS-III and WMS-III which provides two measures of working memory. The potentially confounding presence of Arithmetic combined with the redundancy of assessing WMI on both the WAIS-III and WMS-III would support consideration of condensing this complex structure into a single, stable measure of attentional ability.

Using composite algorithms, the confounded subtest may be removed and Spatial Span (SSp), administered as part of the WMS-III, may be substituted both to increase testing efficiency and broaden the range of attentional measures by inclusion of a non-verbal task. This slight modification has the added advantage of

congruence with existing WAIS-III and WMS-III scores allowing subsequent analyses to continue (i.e., percentiles, discrepancy analyses) within the conventional WAIS-III framework.

Using the computerised RAPT spreadsheet, look-up tables may be produced using the normative information provided in Table 8.2. The requisite normative data for Spatial Span (SSp), Digit Span (DSp) and Letter-Number Sequencing (LNS) are obtained from the WAIS-III/WMS-III technical manual. Intercorrelation between DSp and SSp were obtained by averaging the intercorrelations for three age groups provided in the technical manual (.42+.42+.37)/3 =0.40; Wechsler, 1997a).

Table 8. 2

Example One Scaled Scores and Normative Data

Subtest Norms			
1	DSp	M =	10
		SD =	3
		r =	.90
2	LNS	M =	12
		SD =	3
		r =	.82
3	SSp	M =	11
		SD =	3
		r =	.79

Intercorrelations			
		LNS	SSp
	DSp	.57	.40
	LNS		.45

Normative information is entered into the "Three Subtest Composite" worksheet of the RAPT composite calculator Microsoft Excel spreadsheet producing the output provided in Figure 8.1 below. Areas highlighted in green in the programme indicate required data that must be provided, in this case the scaled score means, standard deviations, variances, reliabilities and intercorrelations.

Figure 8. 1 Three subtest composite calculator for an alternative WM composite.

Three Subtest Composite

NORMS:

	mean	Sd	var	reliability	n
DSp	10	3	9	0.9	3
LNS	10	3	9	0.82	
SSp	10	3	9	0.79	

INTERCORRELATIONS:

	LNS	SSp
DSp	0.57	0.4
LNS		0.45

COVARIANCE :

	LNS	SSp
DSp	10.26	7.2
LNS		8.1

COMPOSITE MEAN

30

COMPOSITE SD

7.24982758

SUM rkk	2.51
2*SUM rk(k-1)	2.84

COMPOSITE RELIABILITY

0.91609589

COMPOSITE SEe

4.15865931

COMPOSITE Sep

6.01438874

As indicated in the figure, the composite calculator provides output (highlighted in yellow) for the composite mean, standard deviation, reliability and standard errors of estimate and prediction. The spreadsheet automatically applies these composite statistics to the formation of a lookup table, as demonstrated in Figure 8.2.

Figure 8.2 provides the deviation quotient, percentile and ninety-percent test and re-test composite intervals for all possible sums of scaled scores. It is notable, that any aspect of these automatic analyses may be modified to suit various clinical requirements. For example, ninety-five percent confidence intervals may be calculated simply by modifying the z-score used in the spreadsheet for confidence interval calculations. The composite calculator spreadsheet is efficient and produces look-up tables which readily facilitate normative interpretations for any given test score combination.

Figure 8. 2 Three subtest composite look-up tables for an alternative WM composite

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
3	44	0.0%	42	56	39	59
4	46	0.0%	44	58	41	61
5	48	0.0%	46	59	43	62
6	50	0.0%	48	61	45	64
7	52	0.1%	50	63	47	66
8	54	0.1%	51	65	48	68
9	57	0.2%	53	67	50	70
10	59	0.3%	55	69	52	72
11	61	0%	57	71	54	74
12	63	1%	59	73	56	76
13	65	1%	61	75	58	78
14	67	1%	63	76	60	80
15	69	2%	65	78	62	81
16	71	3%	67	80	64	83
17	73	4%	69	82	65	85
18	75	5%	70	84	67	87
19	77	6%	72	86	69	89
20	79	8%	74	88	71	91
21	81	11%	76	90	73	93
22	83	13%	78	92	75	95
23	86	17%	80	94	77	97
24	88	20%	82	95	79	98
25	90	25%	84	97	81	100
26	92	29%	86	99	83	102
27	94	34%	87	101	84	104
28	96	39%	89	103	86	106
29	98	45%	91	105	88	108
30	100	50%	93	107	90	110

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
31	102	55%	95	109	92	112
32	104	61%	97	111	94	114
33	106	66%	99	113	96	116
34	108	71%	101	114	98	117
35	110	75%	103	116	100	119
36	112	80%	105	118	102	121
37	114	83%	106	120	103	123
38	117	87%	108	122	105	125
39	119	89%	110	124	107	127
40	121	92%	112	126	109	129
41	123	94%	114	128	111	131
42	125	95%	116	130	113	133
43	127	96%	118	131	115	135
44	129	97%	120	133	117	136
45	131	98%	122	135	119	138
46	133	99%	124	137	120	140
47	135	99%	125	139	122	142
48	137	99%	127	141	124	144
49	139	100%	129	143	126	146
50	141	99.7%	131	145	128	148
51	143	99.8%	133	147	130	150
52	146	99.9%	135	149	132	152
53	148	99.9%	137	150	134	153
54	150	>99.9%	139	152	136	155
55	152	>99.9%	141	154	138	157
56	154	>99.9%	142	156	139	159
57	156	>99.9%	144	158	141	161

To analyse a specific clinical case clinicians must calculate the sum of scaled scores and access the relevant deviation quotient and intervals from the look-up tables. For example, the sum of scaled scores on the alternative working memory composite for a clinical case achieving a scaled score of 10 for DSp, 12 for LNS and 11 for SSp may be calculated using formula 5.2 ($X_c = SS_1 + SS_2 + ... + SS_k$), as follows:

$$X_{WMC} = 10 + 12 + 11$$

$$= 33$$

From figure 8.2, this sum of scaled scores is associated with a WM composite standard score of 106, which falls at the 66th percentile, with a ninety-percent confidence interval of 99 and 113 and a ninety-percent re-test confidence interval of 96 and 116. This client's attentional abilities are equal to or better than sixty-six percent of the normal population and the clinician may be ninety-percent confident that this client's true score falls between 99 and 113 and that, at re-test, any score falling between 96 and 116 would not constitute a systematic deviation in working memory performance on the current composite score.

Perhaps most importantly, the reliability of the alternative composite indicated in figure 8.1 is .91. This is a .03 reduction from the WAIS-III WMI reliability coefficient of .94. Substitution of the more clinically meaningful test produces only a minimal reduction in reliability, avoids confounds due to specific dyscalculia and allows the clinician to increase the efficiency of testing by allowing a single measure of working memory to be calculated. Using this information, clinicians might choose to use the alternative working memory composite for this clinical case.

In this example, RAPT methodology is used to investigate the consequences of modifying an existing composite structure in terms of reliability. Such

modifications may substantially improve the clinical meaningfulness and utility of existing analytical structures. Using RAPT methodology, modified composites are not formed at the expense of actuarial interpretive methods and the current example has succinctly demonstrated the use of RAPT algorithms to produce the same level of interpretive data, in terms of indices of error and confidence intervals, as provided in interpretive manuals. Neither reliability nor analytical capacity are substantially compromised by substitution of a more meaningful subtest, a capacity with far reaching implications for practicing clinicians who must modify test batteries to accommodate a myriad of practical factors.

8.3. Investigating Alternative Battery Structure based on Validity Research

At a more thorough level of application, RAPT methodology may be used to investigate the psychometric impact of using alternative cognitive structures to combine and interpret tests. In the following example research on the WMS-III suggests several alternative interpretive structures for the measurement of memory functioning using WMS-III subtests. The psychometric characteristics of various alternative memory composites are investigated in this example using RAPT methodology.

Examination of factor analysis research may suggest that the existing WMS-III analytical structure (i.e., outlined in Table 8.3) is insufficient to express memory performance and clinicians should consider memory subtest combinations other than those provided in the WMS-III manual (Wechsler, 1997c). Implicit in this suggestion is the aim of increasing testing efficiency by eliminating the administration of redundant memory measures. Factor analyses regarding memory structure would suggest that the interpretive composites customarily used in WMS-III possess some redundancy when applied to the task of developing a

comprehensive understanding of an individual's memory functioning. As indicated in Table 8.3, WMS-III memory composites make the distinction between both visual and verbal memory, and immediate and delayed memory, as well as measuring working memory and a unique summary score composed of recognition tasks requiring verbal responses.

As Table 8.3 outlines, the first level in the hierarchy of the WMS-III analytical structure makes use of an overall immediate memory composite (IM) comprised of both visual and verbal recall tasks where memory is evaluated with no delay following presentation of the stimuli, an overall delayed memory composite (GM) comprised of both visual and verbal measures where 30 to 40 minutes has elapsed between presentation of the stimuli and their subsequent free recall, cued recall or recognition and a composite of working memory (WMI). At the second level in the hierarchy of analysis, distinction is made between composites comprised of auditory immediate (AI), visual immediate (VI), auditory delayed (AD), visual delayed (VD) and delayed auditory recognition (ARD) measures. For the purposes of the following discussion both WM and ARD will be disregarded. Closer consideration of the research literature, in fact produces several conflicting interpretations of this structure which will be discussed in terms of their implications for WMS-III test combinations.

Table 8. 3

WMS-III Index Structure

IM	**GM**	**WM**
Logical Memory I (LMI)	Logical Memory II (LMII)	Letter-Number Sequencing (LNS)
Verbal Paired Associates I (VPA I)	Verbal Paired Associates II (VPA II)	Spatial Span (SSp)
Family Pictures I (FPI)	Family Pictures II (FPII)	
Faces I (FI)	Faces II (FII)	
	Logical Memory II Recognition (LMII R)	
	Verbal Paired Associates II Recognition (VPAII R)	

AI	**VI**	**AD**	**VD**	**ARD**
Logical Memory I (LMI)	Family Pictures I (FPI)	Logical Memory II (LMII)	Family Pictures II (FPII)	Logical Memory II Recognition (LMII R)
Verbal Paired Associates I (VPA I)	Faces I (FI)	Verbal Paired Associates II (VPA II)	Faces II (FII)	Verbal Paired Associates II Recognition (VPAII R)

Studies have consistently failed to support distinct measures of immediate and delayed memory on the WMS-III (Millis, Malina, Bowers & Ricker, 1999; Price, Tulsky, Millis & Weiss, 2002; Tulsky, Chiaravalloti, Palmer & Chelune, 2003; Tulsky & Price, 2003; Tulsky, Ivnik, Price & Wilkins, 2003). Thus, despite the prevalence with which clinicians use the five factor structure indicated in Table 8.3, much research suggests that tests and indices of delayed memory provide no unique information to the clinician. If tests which measure memory after a time delay add little unique information to memory measurement then they may be omitted from administration of the WMS-III and clinicians could administer four immediate memory subtests (i.e., LMI, VPAI, FI and FPI) and analyse only the existing WMS-III immediate memory index.

These same studies, on the other hand, provide consistent evidence for distinctions between verbal and visual memory measures (Tulsky, Chiaravalloti, Palmer & Chelune, 2003; Tulsky & Price, 2003; Tulsky, Ivnik, Price & Wilkins, 2003) resulting in Tulsky and his colleagues suggesting a factor analytically based distinction between verbal and visual subtests on the WMS-III. This research seems to suggest the need for memory measurement composites which retain the distinction between input and output modalities (verbal versus visual) but dispense with the longstanding distinction between immediate and delayed recall. Administering only immediate memory measures and retaining a distinction between verbal and visual performance would substantially simplify the analytical structure required for meaningful interpretation of the WMS-III. Specifically, clinicians could administer two tests of verbal immediate memory to calculate the auditory immediate index (AI) and two tests of visual immediate memory to calculate the visual immediate index (VI). This structure is outlined in Table 8.4.

Table 8. 4

Model 1a for WMS-III Subtest Combination

Memory 1a
Logical Memory I (LMI)
Verbal Paired Associates I (VPA I)
Family Pictures I (FPI)
Faces I (FI)

Auditory Memory 1a	**Visual Memory 1a**
Logical Memory I (LMI)	Family Pictures I (FPI)
Verbal Paired Associates I (VPA I)	Faces I (FI)

There is some concern, however, regarding the efficacy of the standard visual memory measures on the WMS-III, specifically the subtests requiring recognition of faces (Faces I and II). Several studies demonstrate that recall of faces requires a specific type of visuospatial processing (Farah, Wilson, Drain & Tanaka, 1998; McCarthy & Warrington, 1990). In addition the WMS-III Faces subtests rely on recognition, rather than the free recall methodology employed in other WMS-III visual memory subtests (Hawkins & Tulsky, 2004). Finally, validity studies using confirmatory factor analytic techniques demonstrate that the Faces subtests have insufficient communality with other tests loading on visual memory factors (Millis, Malina, Bowers, & Ricker, 1999; Tulsky & Price, 2003; Wilde et. al. 2003). The uniqueness of the Faces subtests is further confirmed by very low intercorrelations with other WMS-III subtests. For example, the correlation between Faces I and Logical Memory I is only .14. Perhaps more importantly, Faces I correlates only .3 with Family Pictures I (Wechsler, 1997a). In fact, the omission of the Family Pictures and Faces measures from the forthcoming WMS-IV suggests that the

dissatisfaction in these measures has penetrated to the level of the test publisher. A memory composite which substitutes the Visual Reproduction subtests for Faces (i.e., LMI, VPAI, VRI and FPI) may provide a more suitable compromise. This is expressed in Table 8.5 below as Model 1b.

Table 8. 5

Model 1b for WMS-III Subtest Combination

Memory 1b	
Logical Memory I (LMI)	
Verbal Paired Associates I (VPA I)	
Family Pictures I (FPI)	
Visual Reproduction I (VRI)	
Auditory Memory 1b	**Visual Memory 1b**
Logical Memory I (LMI)	Family Pictures I (FPI)
Verbal Paired Associates I (VPA I)	Visual Reproduction I (VRI)

Regardless of construct validity research clinicians remain wary of under-eliciting behaviours which they believe to be clinically meaningful. Some may persist in believing that there are individuals for whom specific amnestic disorders can be best detected only through delayed recall measures. In fact, as discussed in chapter four, there may be some sense to retaining a dissociation which appears to be clinically meaningful, as factor analytic research indicates some support for the dissociation of specific cognitive factors in specific clinical samples. In the case of delayed and immediate memory, for example, Delis and colleagues (2003) demonstrated such dissociations in samples of patients with Alzheimer's dementia. This may indicate that the parsimonious structure indicated in previous examples

would fail to elicit meaningful memory performances for some clients. This is a compromise which is perhaps not warranted by increased efficiency.

To accommodate for all research-based concerns an alternative memory composite comprised of Logical Memory I, Logical Memory II, Visual Reproduction I and Visual Reproduction II may meet both clinical conservatism and psychometric rigour and is summarised in Table 8.6 as Model 2. Such a composite removes the potential confound of the most commonly employed visual measures (Faces and Family Pictures) and retains the possible information implicit in both immediate and delayed memory subtests.

Table 8. 6

Model 2 for WMS-III Subtest Combination

Memory 2	
Logical Memory I (LMI)	
Logical Memory II (LMII)	
Visual Reproduction I (VRII)	
Visual Reproduction II (VRII)	
Auditory Memory 2	**Visual Memory 2**
Logical Memory I (LMI)	Visual Reproduction I (VRI)
Logical Memory II (LMII)	Visual Reproduction II (VRII)

The composites outlined in model two deliberately combine immediate and delayed measures. This is suggested by two opposing rationales. First, if immediate and delayed memory tests measure in effect the same construct, as is suggested by failure to dissociate in factor analyses, then they can be considered essentially as replications that will be more highly correlated than other measures. Arguably

delayed recall could be just as effective as immediate memory in measuring the cognitive construct of memory and their combination is likely to produce interpretable and reliable composites. Alternately, if delayed recall is in fact more sensitive to specific memory deficits, then delayed measures are in principle a better reflection of what is meant by "memory", that is the capacity to retain and recall information over time. Of course, standardised administration of WMS-III delayed measures necessitates administration of immediate recall which may be added to the composite to increase stability, based on the potentially strong intercorrelations between immediate and delayed responses to the same stimuli.

The choice of either of these models may be further informed by consideration of the second psychometric characteristic, reliability. Clinicians wishing to use this research literature to directly inform clinical practice may wish to evaluate the consequences of these different memory composites. Information regarding the reliability of such combinations may provide additional weight to a particular argument and, at least, would provide clinicians with the data necessary to defend a particular choice. Table 8.7 below presents the reliability coefficients for the overall memory composites for Model 1a, Model 1b and Model 2. Calculations are again computed using the composite calculator spreadsheet, the output and look-up table for which are provided in full in Appendix H.

As indicated in Table 8.7, the alternative subtest combinations of Model 1b and Model 2 both constitute an improvement in reliability above that provided by Model 1a (i.e., the existing IM composite from the WMS-III). Reliability is improved most by the substitution of the Visual Reproduction I subtest for Faces I in Model 1b. However, the alternative memory composite in Model 2 combining delayed and immediate trials for only LM and VR is almost equally as reliable.

Table 8. 7

Reliability and Standard Error Coefficients for Models 1a, 1b and 2

	Reliability	**SE_E**	**SE_P**
Model 1a Memory Composite	.914	4.20	6.08
Model 1b Memory Composite	.935	3.71	5.34
Model 2 Memory Composite	.923	4.00	5.78

These results indicate that clinicians would not risk reducing reliability by using these alternative memory composites and in fact would somewhat improve the stability of memory measurement.

The test combinations can also be investigated in terms of testing efficiency, as indicated in Table 8.8. Administration times reported in this are derived from average administrations provided by Axelrod (2001). Visual Reproduction estimates are based on the assumption that this subtest takes no longer than the Family Pictures subtest.

Table 8. 8

Administration Time for Models 1a, 1b and 2

	Administration Time
Model 1a Memory Composite	20.6 minutes
Model 1b Memory Composite	21.9 minutes
Model 2 Memory Composite	17.4 minutes

As indicated in the table above, a combination of the two Logical Memory trials with the two Visual Reproduction trials (Model 2) is the most efficient, taking just over seventeen minutes to administer. In comparison the subtests included in the existing WMS-III immediate memory index (Model 1a) takes just over twenty minutes, while the substitution of Visual Reproduction for Faces takes almost an additional two minutes (Model 1b). These results highlight that clinicians could choose the model most indicated by validity research (Model 2) with the added benefit of a slight increase in reliability and a slight decrease in administration time.

Regardless of which model is chosen, however, RAPT methodology provides clinicians with concrete psychometric data to justify recombination of WMS-III subtests. This may be useful to clinicians who risk criticism from various sources when they need to modify "fixed" battery structures. Using RAPT methodology, clinicians could use any alternative structure, with empirical data regarding the impact in terms of reliability. As an added advantage, use of the look-up tables calculated using computerised RAPT methodology facilitate the same level of analytical structure and convenience available using existing fixed batteries, such as the WMS-III and WAIS-III.

In the current example, the choice of either of the more reliable alternative memory composites is not achieved at the expense of analytical methodology. Using RAPT, clinicians can modify the testing battery to achieve optimal validity and clinical utility for a particular client or situation, while still capitalising on specific knowledge regarding reliability and actuarial methodologies.

8.4 Investigate Alternative Battery Structure based on Cognitive Theory

RAPT methodology may also be used to investigate the psychometric impact of using alternative cognitive structures based on cognitive theories, rather than factor analytically determined constructs. As discussed in chapter three, the cognitive domains most familiar to clinicians are in part a function of the original tests devised by theoreticians. While clinicians tend to find investigation of constructs such as processing speed, verbal functioning or working memory meaningful, alternative theories of cognitive functioning may provide informative data, to clinicians, which is unavailable in standardised analytical structures. RAPT methodology facilitates the investigation of the psychometric properties of such constructs as well as their use in clinical inference.

As discussed in section 6.2.1, Cattell and Horn's model of fluid and crystallised intelligence may be used to contrast education-related activities (crystallised intelligence; G*c*) with speeded and analytical tasks (fluid intelligence; G*f*) to provide information regarding age-related changes in cognition (Carroll, 1993; Cattell & Horn, 1978; Horn, 1988, 1998; Horn & Cattell, 1966; Stankov & Horn, 1980). Kaufman and Lichtenberger (2002) have mapped WAIS-III subtests into a five factor model based on the expanded G*f*-G*c* model, as indicated in Table 8.9.

Table 8. 9

Gf-Gc Model of WAIS-III Subtest Combination.

Crystallized Intelligence	**Fluid Intelligence**	**Broad Visualization**	**Short-Term Memory**	**Broad Speediness**
Information	Matrix Reasoning	Picture Completion	Arithmetic	Digit Symbol-Coding
Vocabulary	Block Design	Block Design	Digit Span	Symbol Search
Comprehension	Object Assembly	Object Assembly	Letter-Number Sequencing	Object Assembly
Similarities	Similarities	Matrix Reasoning		
Picture Arrangement	Picture Arrangement			
	Arithmetic			

(Derived from Kaufman & Lichtenberger, 2002)

The test groupings indicated in this table may be investigated using RAPT methodology. Kaufman and Lichtenberger (2002) provide norm-based algorithms with which clinicians may analyse and interpret the alternative cognitive composites. These algorithms do not acknowledge the individual contribution of each measure to the construct in terms of intercorrelations. Additionally, while these authors provide reliability coefficients and indices of standard error (SE_M), they provide no methodology by which clinicians could calculate these psychometrics if clinicians must modify the alternative composites. In fact, Kaufman and Lichtenberger (2002) provide an entirely fixed analytical structure based on the WAIS-III normative sample and while clinicians may use the composite algorithms provided in this publication they are constrained to the exact composites for which the authors provide conversions. In contrast, RAPT methods may be used to provide the same level and type of alternative composites with the added advantage of complete flexibility.

As demonstrated above, the reliability of this alternative structure can be informed through use of RAPT methodology to calculate composite reliability coefficients. Table 8.10 below presents the reliability coefficients for the five G*f*-G*c* composites based on WAIS-III subtests. Calculations are computed using the composite calculator spreadsheet, the output and look-up table for which are provided in full in appendix I.

Table 8. 10

Reliability and Standard Error Coefficients for Gf-Gc composites

	Reliability	**SE_E**	**SE_P**
Crystallised Intelligence	.960	2.94	4.21
Fluid Intelligence	.955	3.09	4.43
Broad Visualization	.932	3.78	5.44
Short-Term Memory	.936	3.66	5.27
Broad Speediness	.883	4.82	7.04

As indicated in Table 8.10, the alternative combinations of WAIS-III subtests demonstrate robust stability. Such results indicate that combination of WAIS-III subtests based on the G*f*-G*c* model using RAPT methodology would:

a) Provide composite scores which are at least as interpretable as those based on the existing WAIS-III analytical structure;

b) Provide estimates of composite reliability, unavailable using the norm-based algorithms provided in Kaufman and Lichtenberger (2002), which indicate psychometric stability, and;

c) Facilitate ease of clinical use, based on look-up tables (provided in appendix I) similar to those provided in the WAIS-III administration and scoring manual (Wechsler, 1997b) and readily produced using programmed spreadsheets.

8.5 Structuring and Analysing a Clinical Battery

At the highest level of influence, RAPT methodology can be used to structure an entire battery of measures specifically suited to a clinical situation. The following example will return to the PCA, reported in section 6.2.1 and discussed in section 6.2.6, to demonstrate the application of RAPT methodology to the structure and analysis of an entire clinical battery. The aim is to demonstrate how clinicians could improve the psychometric robustness, reduce the complexity and increase efficiency of clinical batteries, based on a consideration of construct validity, clinical factors and use of composite methodology. In the current example, evaluation of validity, determination of reliability and consideration of clinical meaningfulness will be undertaken to produce clinical composites from an available test pool used by a clinical practice in Australia. The resulting test combinations will be used to analyse a clinical case, again using analytical tools developed using computerised RAPT methodology.

8.5.1 Review of Principle Components Analysis

As presented in section 6.2, a principal components analysis (PCA) with promax rotation was conducted on a normative sample (n = 1045). The KMO measure of sampling adequacy was .81 indicating that the data was highly factorable and five components with eigenvalues greater than one, were retained. The subtests which comprised each component are summarised in Table 8.11 below.

Table 8. 11

Components derived from Principle Components Analysis of an Australian Normal Sample

Component	Subtests
Word Knowledge (WK)	WAIS3-IN, BNT, WAIS3-VO, WAIS3-CO, WAIS3-SI, STW, WRAT3-Reading, BNT
Processing Speed (PS)	SDMT-Written and Oral, WAIS3-SS, TMT-Part A and B, WAIS3-DSy
Stroop (ST)	STROOP-Colour, Word and Colour/Word
Working Memory (WM)	WAIS3-LNS, WAIS3-DSp, WAIS3-AR
Verbal Fluency (VF)	COWAT, ANIMALS

The measures included in this PCA reflect those commonly used by clinicians in a particular clinical practice in Australia. This PCA was used to evaluate the structure present within a pool of measures, which are drawn upon to construct semi-flexible batteries as dictated by individual client needs. As has been discussed, the practice of selecting a semi-flexible battery from a limited pool of possible measures is typical practice for assessment clinicians and the battery selection methodology reflected in this example is highly typical of how assessment clinicians function.

As discussed in chapter three, examination of the constructs which are empirically demonstrated within a particular test pool and client demographic

constitutes best-practice in terms of construct validity. The group of normally functioning adults analysed in the PCA share similar demographics, in terms of ethnicity and culture, to typical clients of the practice in question. In light of the potential impact of these demographics on performance of cognitive tests, conducting validity research on a representative normative sample provides distinct advantages for clinicians whose clients are drawn from other than North American samples. Based on the results of PCA, measures may be selected to form composites which have been empirically validated on individuals who are likely to be demographically similar to clients.

As previously discussed, using the results of validation studies to produce interpretable composites is a complex process which requires careful balancing between reliability, validity, and clinical issues:

a) In considering possible test combinations composite reliability is important;
b) This should not, however, take precedence over consideration of validity;
c) Neither psychometric consideration should compromise clinical utility, as the combination of practically unsuitable tests is highly likely to produce uninterpretable composite scores.

The structure of each composite, suggested by the PCA, is discussed in the following sections exemplifying attempts to balance these issues in the formation of a clinical battery.

8.5.2 Word Knowledge (WK)

As discussed in section 6.2.6, tests which load on the WK component do not contribute equally to variance within the domain and may be ranked in terms of

factor loadings. On the other hand, subtests are not equally reliable or versatile. In section 6.2.6, RAPT methodology was used to determine the composite reliability associated with WK composites based on various test combinations. In this example, subtests which are likely to be most useful to clinicians in the practice referred to will be chosen.

For example, these clinicians tend not to administer the WAIS-III Comprehension subtest which is a lengthy measure, is the second least reliable subtest, and contributes only to the construction of IQ scores. The WAIS-III Information subtest is omitted due to its vulnerability to formal education and the Boston Naming Test is omitted due to its particular sensitivity to pathology. Inclusion of the BNT could mask deficits and as it is the least reliable subtest this omission is also justified in terms of reliability. The remaining subtests are outlined in Table 8.12.

Table 8. 12

Word Knowledge Subtests

Test	Component Loading	Reliability Coefficient
WAIS3-VO	.801	.93*
WAIS3-SI	.711	.86*
STW	.655	.83**
WRAT3-Reading	.650	.95***

* IN, VO, CO and SI reliability coefficients based on Wechsler (1997a).

** STW reliability coefficient based Baddeley, Emslie & Nimmo-Smith (1992).

***WRAT-3 Reading reliability coefficient based on Wilkinson (1993).

For the purposes of demonstration the word knowledge composite will combine verbally articulated measures of vocabulary, verbal abstraction, and reading

ability and a non-verbally articulated measure of vocabulary. This combination has the added advantage of efficiency, as most of these measures are used in other analyses in the battery. Again, the composite calculator was used to develop look-up tables which are provided in conjunction with computations in appendix J.

Rescaling all composite measures to a single scaling system simplifies computations and ensures that each measure contributes proportionally to composite reliability. To compute the basic composite statistics, scaled score means and standard deviation, and reliabilities and intercorrelations from the PCA were used (see Table 8.13 and 8.14).

Table 8. 13

Normative Means, Standard Deviations and Reliabilities for WK Subtests

	Mean	**SD**	$r_{kk'}$
WAIS3-VO	10	3	.93
WAIS3-SI	10	3	.86
STW	10	3	.83
WRAT3-Reading	10	3	.95

Table 8. 14

Normative Intercorrelations between WK Subtests

Subtest	**WAIS3-SI**	**STW**	**WRAT3-Reading**
WAIS3-VO	.591	.534	.666
WAIS3-SI		.381	.423
STW			.611

The word knowledge composite mean, standard deviation, reliability, standard error of estimate, standard error of prediction, ninety-percent confidence intervals and ninety-percent re-test confidence intervals were calculated using the "Four Subtest Composite" worksheet of the composite calculator. The WK composite is defined by the norms outlined in Table 8.15 (see appendix J for the composite calculator and look-up tables).

Table 8. 15

Normative Data for Word Knowledge Composite

	Normative Data
WK Mean	40
WK Standard Deviation	9.68
WK Reliability	.959
WK Standard Error of Estimate	2.99
WK Standard Error of Prediction	4.27

An individual case may be readily analysed using the WKC look up tables provided in appendix J. The sum of scaled scores for an individual obtaining a WRAT-3 Reading scaled score of 8, WAIS-III Vocabulary scaled score of 9, Similarities scaled score of 8 and Spot-the-Word scaled score of 11 is calculated using formula 5.2 ($X_c = SS_1 + SS_2 + ... + SS_k$, as follows:

$$X_{WK} = 8 + 9 + 8 + 11$$

$$X_{WK} = 36$$

The WK sum of scaled scores is then referred to Table J.1 revealing that the client obtained a WKC of 94 indicating that his word knowledge falls within the Average range and is as good as or better than thirty-four percent of the population. The

clinician can be ninety-percent confident that this client's true ability lies between a standard score of 89 and 99 and ninety-percent confident that a retest score of between 87 and 101 would not reflect systematic change.

8.5.3 Processing Speed (PS)

As with the WK composite, choice of measures to be included in the PS composite will be based on consideration of reliability and clinical utility for the practice in question, as well as loadings on the PS component. Reliability and component loadings are indicated in Table 8.16.

Table 8. 16

Processing Speed Subtests

Test	Component Loading	Reliability Coefficient
SDMT-Written	.874	.80*
SDMT-Oral	.807	.76*
WAIS3-SS	.777	.77**
TMT-Part A	-.730	.89***
TMT-Part B	-.682	.92***
WAIS3-DSy	.627	.84**

* SDMT reliability coefficients based on Smith (1991) manual.

** WAIS-3 reliability coefficients based on Wechsler (1997a).

***TMT reliability coefficient based on Charter et. al.(1987).

Based upon reliability alone the Trail Making Test (TMT) and Digit Symbol seem good choices and would contribute to a highly stable composite. Clinically, however, two aspects of these measures must be considered.

First, is that combining two measures requiring graphomotor output, such as the TMT and DSy, with other speeded tests into a processing speed composite would retain the confound of graphomotor ability with the more general concept of processing speed. In terms of clinical utility these tests may be poor choices, a possibility which is compounded by the fact that the DSy and the TMT have the weakest loadings on the component. In comparison, the Symbol Digit Modalities Test (SDMT) can be invaluable in discriminating between a central processing speed problem and impaired graphomotor abilities by comparing written and oral trials. This capacity makes it preferable to Digit Symbol and Symbol Search, which is again compounded by the fact that both SDMT trials load highest on the processing speed component. Inclusion of this measure in the composite is justified in terms of validity and clinical utility.

The second issue relates to the role of TMT. Part B of this measure demonstrates greater vulnerability to pathology than Part A. While both measures are valuable in evaluating difficulties in mental flexibility, in this case they may detract from the uni-dimensional nature of the processing speed composite by adding the element of executive processing. Ideally, another verbal processing speed measure would be preferable from a clinical perspective. In fact, as an aside, the results of these combined analyses may alert clinicians at the practice in questions of the need to source such a measure. However, for the purposes of the current demonstration a combination between a single graphomotor task (TMT A) and two parts of the SDMT is likely to be useful for the practice in question. Normative data for subsequent analyses are provided in Tables 8.17 and 8.18. Again, intercorrelations between subtests are from the same analysis as the PCA and measures are rescaled to a single scaling system. It is notable that standardisation of the TMT Part A includes reversing the sign of the z-score to accommodate for the

fact that in this test an elevated score is a sign of pathology, rather than high scores indicating increasingly superior performance.

Table 8. 17

Normative Means, Standard Deviations and Reliabilities for PS Subtests

	Mean	**SD**	$r_{kk'}$
SDMT-Written	10	3	.80
SDMT-Oral	10	3	.76
TMT-Part A	10	3	.89

Table 8. 18

Normative Intercorrelations for PS Subtests

Subtest	**SDMT-Oral**	**TMT-Part A**
SDMT-Written	.853	.518
SDMT-Oral		.449

The RAPT composite calculator was used to provide look-up tables, based on the normative data in outlined in Table 8.19 (see appendix J for full results).

Table 8. 19

Normative Data for Processing Speed Composite

	Normative Data
PS Mean	30
PS Standard Deviation	7.73
PS Reliability	.917
PS Standard Error of Estimate	4.13
PS Standard Error of Prediction	5.98

Again, use of the look-up tables can be demonstrated with a clinical case. An individual obtaining a SDMT-Written scaled score of 6, SDMT-Oral scaled score of 7 and TMT-A scaled score of 6 can be analysed using the PSC Table J.2. First, the PS sum of scaled scores is calculated using formula 5.2 ($X_c = SS_1 + SS_2 + ... + SS_k$, as follows:

$$X_{PS} = 6 + 7 + 6$$
$$X_{PS} = 19$$

Referring to Table J.2, the client obtained a PSC of 79 based on a sum of scaled scores of 19 indicating that his processing speed falls in the Below Average range and is as good as or better than eight percent of the population. The clinician can be ninety-percent confident that this client's true ability lies between a standard score of 74 and 87 and ninety-percent confident that a retest score of between 71 and 90 would not reflect systematic change.

8.5.4 Verbal Fluency (VF)

Tests included in the PCA have dissociated into a distinct verbal fluency component. As discussed in section 6.2.2, however, combining these measures

reduces the reliability of the overall composite below that of the COWAT alone. However, both measures demonstrate vulnerability to the influence of pathology and in combination may make a potentially important comparison with the Word Knowledge composite. For the purposes of the current example, therefore, a two subtest composite of verbal fluency was calculated based on the normative data provided in Table 8.20.

Table 8. 20

Normative Means, Standard Deviations, Reliabilities and Intercorrelations for VF Subtests

	Mean	**SD**	$r_{kk'}$	**Intercorrelation**
COWAT	10	3	.82*	.402*
ANIMALS	10	3	.55*	

*Intercorrelations and reliability coefficients derived from the PCA sample.

VF lookup tables were calculated using the composite calculation programme. Again, results are provided in appendix J and summarised in Table 8.21 below.

Table 8. 21

Normative Data for Verbal Fluency Composite

	Normative Data
VF Mean	20
VF Standard Deviation	5.02
VF Reliability	.775
VF Standard Error of Estimate	6.26
VF Standard Error of Prediction	9.47

As with WK and PS, the look-up table facilitates analyses based on the summed scaled scores for VF subtests. Therefore, the sum of scaled scores for an individual obtaining a COWAT scaled score of 4 and ANIMALS scaled score of 3 can be calculated using formula 5.2 as follows:

$$X_{VF} = 4 + 3$$
$$X_{VF} = 7$$

Referring to Table J.3, the client obtained a VFC of 61 indicating that his verbal fluency falls within the Extremely Low range and is as good as or better than less than one percent of the population. Given the lower reliability of the composite, however, confidence intervals are broad: the clinician can be ninety-percent confident that this client's true ability lies between a standard score of 60 and 80; and ninety percent confident that a retest score of between 54 and 86 would not reflect systematic change.

8.5.4 Stroop (ST) and Working Memory (WM)

The third component indicated by PCA includes all three components of the Stroop Colour-Word Association Test and the final component consists of the WAIS-III working memory composite. Both of these combinations can be analysed using RAPT methodology or may be analysed following interpretive methodologies indicated in their respective interpretive manuals. For the purpose of brevity, these constructs are not considered further in the current example.

8.5.5. Discrepancy Analyses

In this demonstration, three new composites, WKC, PSC, and VFC, have been generated based on the capacity of RAPT methodology to facilitate normative

analyses. As ipsative comparisons between domains are often equally important to clinicians, these will be demonstrated following. As discussed in chapter five, the probability and frequency of discrepancies between composites may either be computed anew for each comparison or the levels necessary for significance and abnormality can be computed and then compared to the actual differences.

Each of the possible computational methods requires calculation of the correlation between two composites using formula 5.7 and based on the intercorrelations between composite subtests (see Tables 8.22, 8.23 and 8.24) derived from the PCA sample.

Table 8. 22

Intercorrelations between WK Subtests and PS Subtests

Subtest WKC:PSC	SDMT-W	SDMT-O	TMT-A
WRAT3-Reading	.200	.157	.061
WAIS3-VO	.109	.025	.011
WAIS3-SI	.101	.054	.073
STW	.066	.066	.012

Table 8. 23

Intercorrelations between WK Subtests and VF Subtests

Subtest WKC:PSC	COWAT	ANIMALS
WRAT3-Reading	.445	.253
WAIS3-VO	.335	.232
WAIS3-SI	.271	.277
STW	.357	.259

Table 8. 24

Intercorrelations between VF Subtests and PS Subtests

Subtest WKC:PSC	SDMT-W	SDMT-O	TMT-A
COWAT	.159	.173	.176
ANIMALS	.286	.335	.255

As with the statistics and tables used in normative analyses, composite intercorrelations are readily computed using programmed spreadsheets. Again, the worksheet necessary to conduct these calculations is provided as an appendix to the thesis and the results of calculations provided in appendix J. The discrepancies between composites are required for these computations and are provided in Table 8.25.

Table 8. 25

Example Discrepancies between WK, PS and VF Deviation Quotients

Composites	Discrepancy
WK – PS	15
WK – VF	33
PS - VF	18

Following this the significance and abnormalities associated with composite discrepancies were evaluated using the composite calculator. Figures 8.3, 8.44 and 8.5, taken from the spreadsheet output, provide the cut-off scores for significance at $p<.05$ and $p<.01$ and the between-composite discrepancies associated with the most extreme ten, five and one percent of the normal distribution.

The results in figure 8.3 indicate that the 15 point discrepancy between WKC and PSC was significant at p<.01. Differences of this magnitude, however, are expected to occur in approximately 23% of the population. This suggests that while this gentleman demonstrates better word knowledge than processing speed, this is not unusual and is likely to reflect normal range variation. This is not really surprising given the very low correlation between the two composites (r = .11). These two abilities are clearly unrelated and variation in one is unlikely to influence the other. This highlights that just because a comparison can be made, does not mean it should.

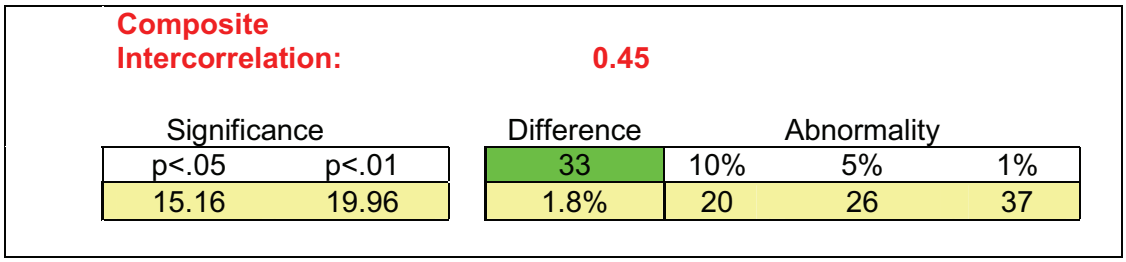
Figure 8. 3 Significance and abnormality for WKC and PSC discrepancy

Composite Intercorrelation: **0.11**

Significance		Difference	Abnormality		
p<.05	p<.01	15	10%	5%	1%
10.35	13.63	22.6%	26	33	46

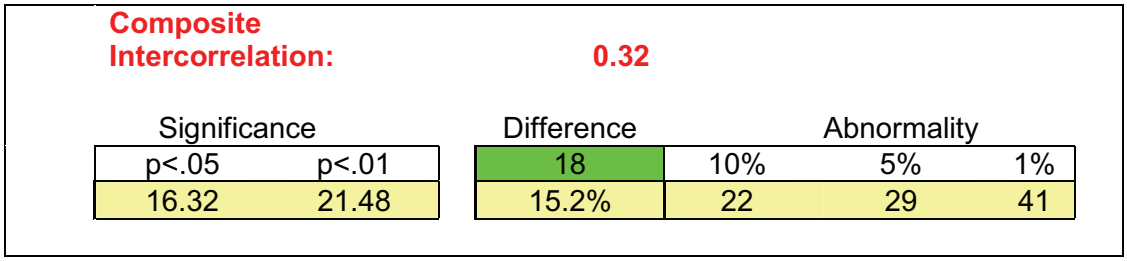
Figure 8.4 considers the discrepancy between the WKC and the VFC. In this instance, the 33 point discrepancy is significant at p <.01. Differences of this magnitude are expected to occur in less than two percent of the population, suggesting that the client's verbal fluency is abnormally low relative to word knowledge: a clear reflection of impairment in this ability. This comparison makes sense in the light of the stronger correlation between the two composites (r = .45) and the fact that each composite evaluates different aspects of verbal processing.

Figure 8. 4 Significance and abnormality for WKC and VFC discrepancy



Finally, figure 8.5 indicates that the 18 point discrepancy between PSC and VFC is also significant at $p < .01$. This comparison is meaningful to clinicians given that both composites measure a different aspect of speeded performance. Differences of this magnitude are likely to occur in approximately fifteen percent of the population, however, and again such a frequency does not constitute abnormality. The results suggest that while the client's processing speed is better than his verbal fluency the difference found between the two composites is not unusual and instead this difference is likely to be reflective of normal range variation.

Figure 8. 5 Significance and abnormality for PSC and VFC discrepancy



The current example has demonstrated the capacity of RAPT methodology to facilitate the structure and analysis of a clinical battery which is based on empirical evaluation of a specific normative population and compiled to service a specific clinical group. As has been extensively discussed, the battery structure employed in the current example may be readily modified when clinicians at the practice in

question need or choose to change their testing battery according to clinical rationale. Regardless of test selection, RAPT methodology readily facilitates use of the most valid, reliable and clinically meaningful collection of test scores for any given client.

8.6 Conclusions

In the current chapter, examples have demonstrated the use of RAPT methodology to address clinical questions. Demonstrations have ranged from the substitution of individual measures into existing composites to the structure of a clinical battery custom-made to suit a particular clinical situation. Throughout the chapter, the use of computerised analyses has demonstrated the ease with which RAPT methodology may be adapted to clinical practice when spreadsheets are used.

Examples have aimed to graphically demonstrate the use of RAPT to integrate the compelling demands of reliability, validity and clinical utility. As demonstrated these three, often conflicting requirements, may be addressed systematically using RAPT methods. Clinicians seeking a solution to the constant complexity of psychometric and clinical requirements may confidently use RAPT methodology to draw clinical inferences which are accompanied by robust psychometric rationales.

At the least invasive level, RAPT facilitates an understanding of working psychometrics and may be useful to teachers of psychometric theory. At the most intrusive level, RAPT provides concrete data regarding the reliability of proposed test combinations and actuarial methodology to conduct both ipsative and normative evaluations. Perhaps most usefully, RAPT provides clinicians with a means of obtaining meaningful and robust inferences when tests from trusted analytical frameworks are compromised for reasons such as client fatigue, time limitations,

confounds associated with specific tests or invalidated protocols. It is hoped that the demonstrations in this chapter provide clinicians with further confidence in the applicability of RAPT to every-day clinical practice based on its capacity to provide analytical structure commonly associated with fixed batteries to a changing collection of cognitive tests accompanied by compelling arguments regarding the psychometric strength of choices.

CHAPTER NINE

GENERAL DISCUSSION, CONCLUSIONS AND IMPLICATIONS FOR CLINICAL PRACTICE

9.1. Overview

Use of a test battery constitutes the norm in assessment practice and allows clinicians to draw inferences regarding the cognitive functioning of individuals in some detail. Psychometric theory indicates that quantifying human behaviours incorporates errors which inevitably occlude understanding of the measured individual's true levels of functioning. While a wealth of literature exists to discuss the control and measurement of error at the level of the individual test, the error associated with analysis and interpretation of a battery of tests is inadequately addressed in clinical practice. Currently clinicians are unlikely to have the necessary information required to make informed estimates of the impact of error on testing batteries. The task of the cognitive assessor is a complex one in which errors must be foreseen and forestalled and the current investigation was driven by the motivation of providing some actuarial methodology by which this difficulty may be addressed by practicing clinicians using a semi-flexible cognitive battery.

The primary goal of the current thesis was to investigate psychometric theory and the typical situations in which cognitive tests are used with the aim of developing a psychometrically and practically driven method of structuring and analysing a semi-flexible battery. The need for some actuarial methodology to serve this role has been increasingly regarded as essential in the clinical literature and has been addressed by several other authors. Psychometric theory was reviewed in this investigation in light of its capacity to directly address issues of measurement error. Additionally, review of clinical practices provided essential information regarding the unavoidable choices which clinicians must make when using cognitive tests on

real individuals. Following this review, the RAPT model proposed in the current thesis was developed to provide clinicians with the necessary data and tools to directly apply psychometric properties of their semi-flexible test batteries in informing and moderating their clinical inferences. RAPT methodology was evaluated using various simulation procedures which examined the efficacy of the proposed methodology to provide essential characteristics of reliability, validity and flexibility to a cognitive battery. This chapter will present a summary of the investigation, consider challenges to the proposed RAPT methodology and indicate directions for future research.

9.2. General Discussion and Conclusion of Results

9.2.1 Applying Reliability, Construct Validity and Clinical Utility to Battery Structure.

The first stage of the thesis was conducted to review the sources of error at the battery level and specifically to examine how reliability, construct validity and pertinent clinical factors impact on the structure and use of a clinically practical and psychometrically robust battery of cognitive tests.

Chapter Two discussed the concept of psychometric reliability and the application of reliability evaluation to the cognitive battery. Reliability was argued to impact directly on clinical decision making and to be underutilised at the level of the cognitive battery. This was found to occur regardless of the ready availability of appropriate methods for reliability estimation both at the level of the individual test and when tests are combined into domain-relevant groupings. It was argued in this chapter that estimation of the internal consistency associated with the measurement of cognitive domains using several tests constituted the most appropriate and practical means of quantifying and reducing random measurement error at the battery

level. Further it was proposed that estimation of internal consistency at the construct level be undertaken as an essential step in battery construction and with due consideration to the strengths and weaknesses of the available reliability coefficients associated with tests.

To the extent that measurement error cannot be estimated using reliability, it must be controlled. Chapter Three discussed the impact of battery structure on controlling errors associated with inference from a cognitive battery. The degree to which "informal" conceptions of cognitive structure direct current battery development was reviewed and the error associated with failure to use an empirically validated structure of domains was discussed. Use of factor analytic techniques to guide battery structure and test selection was reviewed and recommendations for the application of empirically validated structures to semi-flexible battery construction were proposed to accommodate for the strengths and weaknesses of factor analysis-based validation techniques.

It was apparent through review of the psychometric literature in both chapter two and three, that a domain sampling model may be applied to battery construction. Both chapters concluded that tests of specific cognitive constructs may be considered to constitute samples from the same hypothetical domain of construct-specific test items. In terms of improving the measurement of true score (validity) and minimising the impact of random errors across several construct-specific tests (reliability), application of this model would indicate combination of domain-specific tests into a single composite score. In this instance, measurement of a cognitive construct may be accompanied by a precise estimate of reliability allowing the impact of error to directly moderate construct-based clinical inferences.

Of course, psychometric characteristics of tests or batteries are not the primary influence on battery construction which instead is likely to be dictated by

factors of the client, the test setting, the clinician and normative information. Chapter Four, therefore, discussed the impediments to psychometrically driven test selection posed by practical constraints such as client age, education, gender, culture, language, setting factors such as time constraints, clinician factors such as familiarity with measures, and normative issues such as raw-score distributions and scaling.

Chapter four specifically highlighted the necessity of flexible battery construction techniques in accommodating for the vicissitudes of typical clinical practice. Specifically, review of these practical factors demonstrated the fine balance between the need to modify the battery to ensure meaning versus the need for actuarial methodology by which the psychometric characteristics of the battery may be known and employed, and for various clinical decision-making errors to be avoided. In fact, the need for a systematic methodology of structuring and analysing a cognitive battery free from reliance on a fixed battery structure was indicated by this chapter. Chapter four concluded with a call for such a methodology and noted that to successfully accommodate for practical and technical factors in battery construction, such a method must consider reliability and construct validity, as well as clinical constraints.

9.2.2. A Reliable Approach to Psychological Testing.

Chapter Five outlined a reliable approach to psychometric testing (RAPT) which applied psychometric theory and clinical factors to the use of robust battery structure. The RAPT method presented in this chapter focussed on the use of composite scores of domain-specific tests, grouped according to empirically validated domains and moderated by direct estimation of composite reliability. The use of composite scores as a means of formalising validated cognitive structure, controlling domain-specific error and flexibly accommodating for battery

modifications was discussed. Composite methodology, as employed in RAPT, was argued to provide clinicians with concrete data with which to answer common challenges to flexible and semi-flexible battery usage. Specifically, three stages of RAPT methodology were outlined including:

a) The use of valid domains to structure the battery and use of reliability to moderate test selection for each domain;
b) Calculation of composite observed scores, means, standard deviations and reliabilities for each domain and intercorrelations between composite scores for each domain; and,
c) Use of standard score deviation quotients and reliability based indices of error to conduct normative and ipsative analyses.

Following explication of the model, algorithms for the RAPT model were presented. RAPT formulae facilitated calculation of composite scores allowing both normative and ipsative interpretation and based on normative means, standard deviations and reliability coefficients for individual tests and the intercorrelations between tests within the battery.

9.2.3. Examining the Utility of RAPT

For clinicians to deviate from accepted analytical and interpretive practices, evidence of the validity of new practices must be empirically demonstrated. To this end in chapters six, seven and eight RAPT methodology was examined in terms of its capacity to control and measure error at the battery level, its ability to recreate the "known world", and its clinical utility. The results indicated that RAPT provided a suitable method to improve psychometric strengths of cognitive test batteries.

Chapter Six discussed several psychometric and clinical advantages provided by use of composite methodology. Specifically, composite scores such as those employed in RAPT were demonstrated to provide a means of:

a) formalising validated structure;
b) of controlling and understanding composite reliability;
c) of improving the reliability of domain-based scores;
d) of reducing the impact of artifactual errors on test scores and;
e) fundamentally, of providing actuarial methodology with which clinicians could achieve an analytical structure commonly associated with a fixed battery regardless of the flexibility of battery construction.

Chapter Seven evaluated the capacity of RAPT to replicate psychometrically valid and stable battery interpretive structures by replication of the WRAT-4 and WAIS-III summary scores, reliability evaluations and normative data. In this chapter, RAPT was used to re-create WRAT-4 and WAIS-III normative and ipsative tables, summary score intercorrelations and reliability coefficients with a high level of accuracy demonstrating its applicability for clinical practice.

Finally, Chapter Eight applied RAPT to clinical examples to examine the ability of the entire methodology to improve the accuracy, stability and meaningfulness of clinical interpretations. This chapter introduced the use of programmed spreadsheets to facilitate efficient use of RAPT algorithms. RAPT was demonstrated to apply to at four levels of clinical usage:

a) First to investigate the psychometric consequences of test combination;
b) Second to modify existing composites;
c) Third to modify existing battery structure; and,
d) Fourth to structure and analyse a unique and context specific clinical battery.

In this chapter RAPT methods were used to substitute subtests into existing composites to remove a specifically confounded measure, to explore the psychometric impact of alternative memory structure in measurement of memory using WMS-III subsets and finally to develop analytical tools to structure and analyse a setting-specific battery of measures based on the results of a principal components analysis. The demonstration indicated the clinical utility of RAPT methodology and concluded that clinicians could confidently and efficiently employ RAPT when using semi-flexible batteries of cognitive tests in real-world clinical practice. Chapter eight demonstrated the role of the proposed methodology in allowing estimation of reliability, in guiding battery structure, in formalising test combination and ipsative comparisons and in successfully accommodating constraints of the client and test setting.

9.4 Strengths of RAPT Methodology

RAPT has been demonstrated to address several criticisms levelled at semi-flexible batteries. Specifically, RAPT provides a means by which vital psychometric characteristics may be calculated and applied to clinical decision making at the battery level. This provides practicing cognitive assessors with the capacity to moderate the confidence with which they draw clinical inferences directly according to psychometric reliability.

A fundamental weakness of flexible and semi-flexible battery construction is that domain-based inferences are often based solely on informal test combinations with unknown reliability. Implicit in this practice is the presumption that tests provide equally valid and reliable measurement of cognitive domains. Of course, this is not the case and instead individual measures of the same cognitive domains have widely differing reliability and test score meaning. When RAPT methodology

is used to formalise this domain-specific test combination both the individual test reliability coefficients and interrelationships between measures of the same domain are considered in composite analyses.

In fact, RAPT applies psychometric theory with several levels of impact and an important contribution of the current study is the application of actuarial methodology to flexible battery structure. Using RAPT methods, the composite reliability coefficient is directly applied to subsequent normative and ipsative interpretation of composite scores. Clinicians may use RAPT confidence intervals and formal statistical tests of the significance and abnormality of discrepancies between composites. Essentially, RAPT facilitates a level of actuarial methodology typically available in some fixed batteries, such as the WAIS-III, and surpassing that provided by others, such as the HRNB. It is stressed, that a primary contribution of the RAPT method is the readiness with which it may be applied to any collection of measures dictated by the needs of practicing clinicians.

At a fundamental logistical level, RAPT summarises the myriad of complex procedures required to analyse a multi-measure battery into three straight-forward stages of battery usage, the methods for which may be readily programmed into a spreadsheet and used with ease by clinicians. In so doing, RAPT carried the added strength of combining difficult psychometric theory into succinct actuarial processes readily applicable to clinical test usage.

9.5 Limitations and Future Directions

While RAPT methodology is strongly supported by psychometric principles and modelled with a necessary degree of flexibility, the methodology can be immeasurably advanced by future research improving the quality of normative and psychometric information available for individual tests. In fact, this investigation

carries several limitations which dictate the direction for future research and may impede wide-spread application of RAPT methodology.

9.4.1 Norms and Domains are Sample Specific

The inadequacy of current normative information presents a primary barrier to the application of RAPT methodology in clinical practice. While use of clinical composites constitutes an improvement over current best-practice, universal application is likely to be hindered by a dearth of suitable factor analytic research, test intercorrelations, stable reliability coefficients and client-appropriate norms. It should be recognised, however, that these same limitations undermine any application of psychological testing. As has been suggested, ignoring these shortcomings in the clinical literature can in no way be considered to constitute "best-practice"

As discussed, the confidence with which clinicians may apply factor analytic research must be moderated by consideration of the vulnerabilities of this scientific method. Clinicians may be most confident when basing battery structure on domains indicated in analyses of samples which are highly similar to the individual client. This unavoidable fact moderates the degree to which validity information is used in RAPT. As discussed, the individual components of composite scores may be weighted according to the precise degree to which they contribute to domain measurement indicated, for example, by component or factor loadings. While RAPT directly considers reliability which is likely to be readily available, test weightings are not advised in RAPT procedures: a potential weakness which was considered to be better than the error likely when domain-weightings are inappropriately applied. Test weightings derived from an inappropriate factor analysis may encourage a false level of confidence in composite validity. It was considered preferable that clinicians

acknowledge reduced validity and apply this admission to directly moderate clinical inferences.

In fact, this caution holds true for all necessary RAPT statistics, including individual test intercorrelations, reliability coefficients and normative means and standard deviations, all of which are fundamentally sample specific. The capacity of RAPT methodology to accurately represent the true reliability and validity associated with test combination is entirely vulnerable to the availability of appropriate normative information. As normative information is increasingly inappropriate, indices of error produced using RAPT formulae will carry increasing inaccuracies. Fortunately, this is not specific to RAPT but rather is a factor of most CTT based test score analyses and in fact presents a fundamental challenge to the quantification of cognitive behaviours regardless of methodology.

On a less complex, but potentially more influential note, RAPT is specifically confined to those batteries for which normative intercorrelations are available. While means, standard deviations, reliability coefficients and at least some validity data are readily available for many, if not most, cognitive tests, intercorrelations between tests are less frequently available. This specifically impacts on evaluation of the intercorrelations between composites for which test intercorrelations are required. While intercorrelations between domain-specific tests are more readily available, lack of these prevents use of RAPT methodology which relies fundamentally on the correlational relationships between measures. This is a primary limitation of RAPT which will only be alleviated by increasing availability of these essential data.

A possible solution requires practicing clinicians to develop such normative and psychometric data from their own clinical setting. As discussed, the development of normative data specific to a particular clinical practice could provide

the most appropriate means of improving the psychometric quality of battery-based inferences. Such a procedure is logistically difficult and time consuming, however, and the most workable solution to this difficulty may come from the application of modern test theories (e.g., Item Response Theory) to the development of normative test data where CTT fails. A study by Bechger, Maris, Verstralen & Beguin (2003), for example, established IRT as a mathematical extension of CTT and demonstrated the development and use of IRT models to produce CTT properties, such as reliability coefficients, from limited sampling. Specifically, the article provides general formulae which facilitate derivation of reliability coefficients using IRT models. If applied clinically, such research would greatly facilitate application of RAPT.

9.4.2. Limitations due to Classical Test Theory

Composite methodology and the theory of error applied in this thesis are based solely on the assumptions and theorems of Classical Test Theory (CTT; Lord & Novick, 1968; Gulliksen, 1950). As such, it is vulnerable to challenges to CTT present in the current psychometric literature notably the sample-reliance of CTT indices. Nothing prevents application of new theories of measurement error and reliability (e.g., generalisability theory or Item Response Theory) to RAPT, however, which would readily adapt to the formation of composites based on latent traits and evaluated through use of alternative coefficients of score stability.

9.4.3. Limitations due to Complexity of the Methodology

Finally, the time required to undertake complex composite calculations would undoubtedly prove prohibitive to practicing clinicians. The time required to incorporate RAPT methods into analytical techniques would therefore present

another challenge to the clinical application of RAPT. In fact, clinicians who intend to use the methodology at the more involved levels would benefit from the development of computerised scoring software based on proposed structure and composite algorithms. This may be readily accomplished using computer spreadsheets, as demonstrated in Chapter Eight. However, some clinicians may undoubtedly be somewhat daunted by the methodology itself.

9.5 Conclusions

The relationship between the literature and the data that informs clinical practice can be clearly linked to clinical interpretations and decision-making which are at the core of evidence-based practice. Implicit in this phrase, is that clinicians understand the reasons for inferences and how these relate to the data. For example, clinicians speak of reliability as a test characteristic. In fact reliability dictates the weight given to various aspects of clinical data. RAPT is an operationalisation of evidence based practice. The methodology does not dictate that clinicians acknowledge that "reliability matters", but rather that they directly calculate the reliability coefficients associated with the test combinations from which they draw clinical inferences and apply these in decision making. The RAPT methodology ultimately dictates a systematic gathering of psychometric facts associated with the semi-flexible battery and facilitates this by providing the means by which more facts than are ordinarily available may be produced.

Clinicians currently undertake the process of cognitive measurement with widely varying degrees of reference to psychometric theory and the results of this practice, in terms of measurement error, are largely unevaluated. With a wealth of psychometric literature and theory, failure to apply this scientific method at every possible level of measurement is inexcusable leading to justifiable debate regarding

accuracy and stability of the inferences drawn from cognitive batteries. It is true of the current investigation that the questions driving enquiry were not those regarding the "permissibility of statistics", but those pertaining to the drawing of "legitimate inference" (Michell, 1986). To this aim, the RAPT method provides clinicians with the means to develop concrete data regarding the scope and limitations of a flexibly constructed battery. Theory, science and practice are far from successfully melded in current clinical practice (Franzen, 2000), however, it is hoped that RAPT assists clinicians to strive for accurate behavioural quantification based on psychometric science.

REFERENCES

Agranovick, A. V., & Puente, A. E. (2007). Do Russian and American normal adults perform similarly on neuropsychological tests? Preliminary findings on the relationship between culture and test performance. *Archives of Clinical Neuropsychology, 22,* 273-282.

Aiken, L. R. (1991). *Psychological testing and assessment* (7th ed.). Boston: Allyn & Bacon.

Aiken, L. W. (2003). *Psychological Testing and Assessment* (11th ed.). Sydney: Pearson.

Ambrose, P. A. J. (1997). Challenges for mental health service providers: The perspective of managed care organisations. In J. N. Butcher (Ed.). *Personality assessment in managed health care* (pp. 61-72). New York: Oxford University Press.

American Academy of Neurology (AAN) (1996). Assessment: neuropsychological testing of adults: Considerations for neurologists. *Neurology 47*, 592-599.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing International Edition* (7th ed.). Sydney: Prentice Hall.

Arai, H. (2005). Current concept and future research directions of mild cognitive impairment. *Psychogeriatrics 5*, 83-88.

Ardila, A. (1999). A Neuropsychological Approach to Intelligence *Neuropsychology Review, 9*(3), 117-147.

Ardila, A. (2005). Cultural Values Underlying Psychometric Cognitive Testing. *Neuropsychology Review, 15*(4), 185-194.

Ardila, A. (2007). Toward the development of a cross-linguistic naming test. *Archives of Clinical Neuropsychology, 22,* 287-307.

Arkes, H. R. (1981). Impediments to actual clinical judgement and possible ways to minimise their impact. *Journal of Consulting and Clinical Psychology, 49,* 323-330

Atkinson, L. (1991). Some Tables for Statistically Based Interpretation of WAIS-R Factor Scores. *Journal of Consulting and Clinical Psychology, 3*(2)*,* 288-291.

Arnau, R. C., & Thompson, B. (2000). Second-order confirmatory factor analysis of the WAIS-III. *Assessment, 7*(3)*,* 237-246.

Australian Psychological Society (2007). *Code of Ethics.* Victoria: Australian Psychological Society Limited

Axelrod, B. N. (2001). Administration duration for the Wechsler Adult Intelligence Scale-III and Wechsler Memory Scale-III. *Archives of Clinical Neuropsychology, 16*(3), *293-301.*

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1992). *The Speed and Capacity of Language-Processing Test.* Bury St. Edmunds: Thames Valley Test Company.

Baird, A. D. Ford, M., & Podell, K. (2007). Ethnic differences in functional and neuropsychological test performance in older adults. *Archives of Clinical Neuropsychology, 22*(3), 309-318.

Basso, M. R., Carona, F. D., Lowery, N., & Axelrod, B. N. (2002). Practice Effects on the WAIS-III Across 3- and 6- Month Intervals. *The Clinical Neuropsychologist, 16*(1), 57-63.

Basso, M.R., Bornstein, R.A., & Lang, J.M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist, 13,* 283-292.

Bauer, R. M. (2000). The Flexible Battery Approach to Neuropsychological Assessment In R. D. Vanderploeg (Ed.). *Clinician's Guide to*

Neuropsychological Assessment (2nd. ed., pp. 419-448). Mahwah, NJ Lawrence Erlbaum Associates

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Beguin, A. A. (2003). Using Classical Test Theory in Combination with Item Response Theory *Applied Psychological Measurement, 27*(5), 319-334.

Bennett, T. L. (2001). Neuropsychological evaluation in rehabilitation planning and evaluation of functional skills. *Archives of Clinical Neuropsychology, 16*, 237-252.

Bertone, A. Bittinelli, L., & Faubert, J. (2007). The impact of blurred vision on cognitive assessment. *Journal of Clinical and Experimental Neuropsychology, 29*(5), 467-476.

Bialystok, E., & Craik, F. I. M. (2007). Commentary: Bilingualism and naming: Implications for cognitive assessment. *Journal of the International Neuropsychological Society, 33,* 209-211.

Bigler, E. D. (2001). Neuropsychological testing defines the neurobehavioural significance of neuroimaging-identified abnormalities. *Archives of Clinical Neuropsychology, 16*, 227-236.

Bigler, E. D. (2007). A motion to exclude and the 'fixed' versus 'flexible' battery in 'forensic' neuropsychology: Challenges to the practice of clinical neuropsychology. *Archives of Clinical Neuropsychology, 22,* 45-51.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *journal of Neuroscience 17*(1), 352-362.

Boring, E. E. (1923). Intelligence as the tests test it. *New Republic, June*, 35-37.

Bowden, S. C. (2004). The role of factor analysis in construct validity: Is it a myth? *Journal of the International Neuropsychological Society, 10,* 1018-1019.

Bowden, S. C., Carstairs, J. R., & Shores, E. A. (1999). Confirmatory Factor Analysis of Combined Wechsler Adult Intelligence Scale – Revised and Wechsler Memory Scale – Revised Scores in a Healthy Community Sample. *Psychological Assessment, 11*(3), 339-344.

Bowden, S. C., Ritter, A. J., Carstairs, J. R., Shores, E. A. Pead, J. Greeley, J. D., Whelan, G. Long, C. M., & Clifford, C. C. (2001). Factorial Invariance for Combined Wechsler Adult Intelligence Scale-Revised and Wechsler Memory Scale – Revised Scores in a sample of clients with Alcohol Dependency. *The Clinical Neuropsychologist, 15 (1),* 69-80.

Brauer-Boone, K., Victor, T. L., Wen, J., Razani, J., & Ponton, M. (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology, 22,* 355-365.

Brennan, R. L. (1983). *Elements of Generalizability Theory.* Iowa City, IA: American

Brownell, W. A. (1933). On the accuracy with which reliability may be measured by correlating test halves. *Journal of Experimental Education, 1,* 204-215.

Brunner, M., & Sub, H.-M. (2005). Analyzing the Reliability of Multidimentional Measures: An Example from Intelligence Research. *Educational and Psychological Measurement, 65*(2), 227-240.

Busch, R. M., Chelune, G. J. & Suchy, Y. (2005). Using norms in neuropsychological assessment of the elderly. In D. K. Attix & K. A. Welsh-Bohmer (Eds). *Geriatric Neuropsychology* (pp. 133-157). New York: Guilford Press.

Butler, M., Retzlaff, P., & Vanderploeg, R. D. (1991). Neuropsychological Test Usage *Professional Psychology: Research and Practice 22*(6), 510-512.

Byrd, D. A., Jacobs, D. M., Hilton, H. J., Stern, Y., & Manly, J. J. (2005). Sources of errors on visuoperceptual tasks: role of education, literacy, and search strategy. *Brain and Cognition*, *58*(3), 251–257.

Byrd, D. A., Touradji, P., Tang, M.-X., & Manly, J. J. (2004). Cancellation test performance in African American, Hispanic, and White elderly. *Journal of the International Neuropsychological Society*, *10*(3), 401–411.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psycological Test Usage: Implications in Professional Psychology. *Professional Psychology: Research and Practice, 31*(2), 141-154.

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

Cattell, R. B. (1964). Validity and Reliability: A Proposed More Basic Set of Concepts. *Journal of Educational Psychology, 55 (1)*, 1-22.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum.

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement, 15,* 139-164.

Chapman, L., & Chapman, J. (1967). Genesis of popular but erroneous psycho-diagnostic observations. *Journal of Abnormal Psychology, 72*, 193-204.

Charter, R. A., Adkins, T. G., Alekoumbides, A., & Seacat, G. F. (1987). Reliability of the WAIS, WMS, and Reitan Battery: Raw scores and standardization scores corrected for age and education. *The International Journal of Clinical Neuropsychology, 9,* 28-32.

Charter, R. A. (1999). Sample Size Requirements for Precise Estimates of Reliability, Generalizability and Validity Coefficients. *Journal of Clinical and Experimental Neuropsychology, 21*(4), 559-566.

Charter, R. A. (2003). A Breakdown of Reliability Coefficients by Test Type and Reliability Method, and the Clinical Implications of Low Reliability. The Journal of General Psychology, *130*(3), *290-304.*

Charter, R. A., & Feldt, L. S. (2001). Meaning of Reliability in Terms of Correct and Incorrect Clinical Decisions: The Art of Decision Making is Still Alive. *Journal of Clinical and Experimental Neuropsychology, 23*(4), 530-537.

Charter, R. A., & Feldt, L. S. (2002). The Importance of Reliability as it Relates to True Score Confidence Intervals. *Measurement and Evaluation in Counselling and Development, 35,* 104-112.

Chesher, A. (1991). The Effect of Measurement Error. *Biometrika, 78*(3), 4510462.

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment, 6*(4), 284-290.

Cimino, C. R. (2000). Principles of Neuropsychological Interpretation. In R. D. Vanderploeg (Ed.). *Clinician's Guide to Neuropsychological Assessment* (2nd ed.). Mahwah, NJ Lawrence Erlbaum Associates.

Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology, 78*(1), 98-104.

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO PI-R Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Craik, F. I. M., & Lockhard, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behaviour, 11,* 671-684.

Crawford, J. R., & Howell, D. C. (1998). Comparing an Individual's Test Score Against Norms Derived from Small Samples *The Clinical Neuropsychologist, 12*(4), 482-486.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-335.

Cronbach, L. J. (1971). *Test Validation.* In R. L. Thorndike (Ed.). *Educational Measurement (2nd ed. Pp.443-507).* Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*, 281-302.

Cronbach, L. J., & Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha And Successor Procedures. *Educational and Psychological Measurement 64*(3), 391-418.

Cronbach, L. J., Gleser, G. C. Nanda, H., & Rajaratnum, N. (1972). *The dependability of behavioural measures: Theory of generalizability for scores and profiles.* New York: John Wiley.

Crossen, B. (2000). Application of Neuropsychological Assessment Results. In R. D. Vanderploeg (Ed.). *Clinician's Guide to Neuropsychological Assessment* (2nd ed.). Mahwah, NJ Lawrence Erlbaum Associates.

Daubert v. Merrell. (1993). *Daubert v. Merrell Dow Pharmaceuticals*, 113 Supreme Court Reporter (S.CT.) 2786.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95-106.

Delis, D.C., Kaplan, E., & Kramer, J.H. (2001). *The Delis-Kaplan Executive Function System.* San Antonio: The Psychological Corporation.

Delis, D. C., Jacobson, M., Bondi, M. W., Hamilton, J. M., & Salmon, D. P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: Lessons from memory assessment *Journal of the International Neuropsychological Society 9*, 936-946.

Demonet, J.-F., & Thierry, G. (2001). Language and Brain: What is Up? What is Coming Up? *Journal of Clinical and Experiemental Neuropsychology 23*(1), 49-73.

DeVellis, R. F. (2003) *Scale Development: Theory and Applications* (2nd ed.). Thousand Oaks, Calif: Sage Publications.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effects of the Expanded Halstead-Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society, 5,* 346-356.

Donders, J. (2001a). A Survey of Report Writing by Neuropsychologists I: General Characteristics and Content. *The Clinical Neuropsychologist, 15*(2), 137-149.

Donders, J. (2001b). A Survey of Report Writing By Neuropsychologists, II: Test Data, Report Format and Document Length. *The Clinical Neuropsychologist, 15*(2), 150-161.

Dougherty, D. D., Rauch, S. L., & Fischman, A. J. (2004). Positron Emission Tomography and Single Photon Emission Computed Tomography In D. D. Dougherty, S. L. Rauch & J. G. Rosenbaum (Eds.). *Essentials of Neuroimaging for Clinical Practice* (pp. 75-92). Arlington, VA: American Psychiatric Publishing.

Dudeck, F. J. (1979). The Continuing Misinterpretation of the Standard Error of Measurement. *Psychological Bulletin, 86*(2), 335-337.

Dunnette, M. D., & Borman, W. C. (1979). Personnel Selection and Classification Systems. *Annual Review of Psychology, 30*, 477-525.

Echemendia, R. J. (2007). Editorial Introduction to the special edition. *Archives of Clinical Neuropsychology, 22,* 271-272.

Eisman, E.J., Dies, R. R., Finn, S.E., Eyde, L.D., Kay, G.G., Kubiszyn, T.W., Meyer, G. J., & Moreland, K. (2000). Problems and limitations in using psychological assessment in the contemporary and health care delivery system. *Professional Psychology: Research and Practice, 31,* 131-140.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. London: Lawrence Erlbaum Associates.

Fan, X., & Yin, P, (2003). Examinee Characteristics and Score Reliability: An Empirical Investigation. *Educational and Psychological Measurement, 63* (3), 357-368.

Farah, M.J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about faces perception? *Psychological Review, 105,* 482-498.

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed.). Phoenix: Oryx Press.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp.105-146). Phoenix, AZ; Orynx.

Fernandez-Duque, D., & Posner, M. I. (2001). Brain Imaging of Attentional Networks in Normal and Pathological States. *Journal of Clinical and Experimental Neuropsychology, 23*(1), 74-93.

Fischhoff, B. (1975). Hindsight ≠ Foresight: The Effect of Outcome Knowledge on Judgement Under Uncertainty. *Journal of Experimental Psychology, 1*(3), 288-299.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Flynn, J. R. (1994). IQ gains over time. In R. J. Sternberg (Ed.). *The encyclopedia of human intelligence* (pp. 617–623). New York: Macmillan.

Flynn, J. R. (1998a). WAIS-III and WISC-III IQ gains in the United States from 1972 to 1995: How to compensate for obsolete norms. *Perceptual and Motor Skills, 86,* 1231–1239.

Flynn, J. R. (1998b). IQ Gains Over Time: Toward Finding the Causes. In U. Neisser (Ed.), *The Rising Curve: Long-Term Gains in IQ and Related Measures* (pp. 25-66). Washington, DC: American Psychological Association.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54,* 5–20.

Folia, V., & Kosmidis, M. H. (2003). Assessment of Memory Skills in Illiterates: Strategy Differences or Test Artefact? *The Clinical Neuropsychologist, 17*(2), 143-152.

Foster, S. L., & Cone, J. D. (1995). Validity Issues in Clinical Assessment. *Psychological Assessment, 7*(3), 248-260.

Franzen, M. D. (2000). *Reliability and Validity in Neuropsychological Assessment*. New York Kluwer Academic/Plenum Publishers.

Frisoni, G. B., Scheltens, P., Galluzzi, S., Nobili, F. M., Fox, N. C., Robert, P. H., et al. (2003). Neuroimaging tools to rate regional atrophy, subcortical

cerebrovascular disease, and regional cerebral blood flow and metabolism: consensus paper of the EADC. (Review). *Journal of Neurology, Neurosurgery and Psychiatry, 74*(10), 1371-1382.

Gale, S. D., Baxter, L., Connor, D. J., Herring, J. & Comer, J. (2007). Sex differences on the Rey Auditory Verbal Learning Test and the Brief Visuospatial Memory Test-Revised in the elderly: Normative data on 172 participants. *Journal of Clinical and Experimental Neuropsychology, 29*(5), 561-567.

Garb, H. N., & Schramke, C. J. (1996). Judgment Research and Neuropsychological Assessment: A Narrative Review and Meta-Analyses. *Psychological Bulletin, 120*(1), 140-153.

Ghiselli, E. E. (1964). *Theory of Psychological Measurement*. New York: McGraw-Hill.

Gierl, M. J., & ElAtia, S. (2007). Book Review: Adapting Educational and Psychological Tests for Cross-Cultural Assessment. *Applied Psychological Measurement, 31*(1), 74-78.

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.

Goh, H. E. (2006). A new structural summary of the MMPI-2 for evaluating personal injury claimants, PH.D. Thesis, University of Southern Queensland.

Golden, C. J. (1978). *Stroop Colour and Word Test.* Chicago. IL: Stoelting.

Golden, C. J., Purisch, A. D., & Hammeke, T. A. (1988). *Luria-Nebraska Neuropsychological Battery Forms I and II Manual*. Los Angeles: Western Psychological Services.

Goldstein, G. (1997). The Clinical Utility of Standardized or Flexible Battery Approaches to Neuropsychological Assessment In G. Goldstein & T. M. Incagnoli (Eds.). *Contemporary Approaches to Neuropsychological Assessment.* London Plenum Press.

Goldstein, M. A., & Price, B. H. (2004). Magnetic Resonance Imaging In D. D. Dougherty, S. L. Rauch & J. G. Rosenbaum (Eds.). *Essentials of Neuroimaging for Clinical Practice* (pp. 21-74). Arlington, VA American Psychiatric Publishing.

Gollan, T. H., & Fennema-Nostestine, C. (2007). What is it about bilingualism that affects Boston Naming Test Performance? A reply to commentaries. *Journal of the International Neuropsychological Society, 13,* 215-218.

Gollan, T. M., Fennema-Notestine, C., Montoya, R. I., & Jernigan, T. L. (2007). The bilingual effect on Boston Naming Test performance. *Journal of the International Neuropsychological Society, 13,* 197-208.

Goodglass, H., & Kaplan, E. (1983). *Assessment of aphasia and related disorders* (2nd ed.). Philidelphia: Lea & Febiger.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka, & W. F. Velicer (Eds.). *Handbook of psychology: Vol. 2 Research methods in psychology* (pp.143-164). NJ: John Wiley & Sons.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93,* 216-229.

Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice.* Boston: Allyn & Bacon.

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

Guilford, J. P. (1948). Factor analysis in a test-development program. *Psychological Review, 55,* 79-94.

Guilford, J. P. (1950). *Fundamental statistics in psychology and education* (2nd ed.). New York: McGraw-Hill.

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

Guilford, J. P. (1956). *Fundamental Statistics in Psychology and Education* (3rd ed.). London: McGraw-Hill.

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

Guilmette, T. J., Faust, D., Hart, K., & Arkes, H. R. (1990). A National Survey of Psychologists Who Offer Neuropsychological Services. *Archives of Clinical Neuropsychology, 5*, 373-392.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: John Wiley & Sons.

Gulliksen, H. (1987). *Theory of mental tests.* NG: Lawrence Erlbaum.

Halstead, W. C. (1947). *Brain and Intelligence A Quantitative Study of Frontal Lobes* Chicago: The University of Chicago Press.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hawkins, K. A., & Tulsky, D. S. (2004). Replacement of the Faces Subtest by Visual Reproduction within Wechsler Memory Scale - Third Edition (WMS-III) Visual Memory Indexes: Implications for Discrepancy Analysis. *Journal of Clinical and Experimental Neuropsychology, 26*(4), 498-510.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased Judgments of Past Events After the Outcomes Are Known. *Psychological Bulletin, 107*(3), 311-327.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment 7*(3), 238-347.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an Expanded Halstead-Reitan Battery Demographic Corrections, Research Findings, and Clinical Applications*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Taylor, M., & Manly, J. (2003). Demographic effects and demographically corrected norms with the WAIS-III and WMS-III. In D. Tulsky, R. K. Heaton, G. J. Chelune, I. Ivnik, R. A. Bornstein, A. Priftera, & M. Ledbetter (Eds.). *Clinical interpretation of the WAIS-III and WMS-III* (pp.181-210). San Diego, CA: Academic Press.

Henson, R. K., & Roberts, J. K. (2006). Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement, 66*(3), 393-416.

Herlitz, A., Airaksinen, E., & Nordstom, E. (1999). Sex differences in episodic memory: The impact of verbal and visuospatial ability. *Neuropsychology, 13,* 590-597.

Hill-Briggs, F., Dial, J. G. Morere, D. A., & Joyce, A. (2007). Neuropsychological assessment of persons with physical disability, visual impairment or blindness, and hearing impairment or deafness. *Archives of Clinical Neuropsychology, 22,* 389-404.

Hiscock, M. (2007). The Flynn effect and its relevance to neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 296*(5), 514-529.

Hogan, T. P. & Agnello, J. (2004). An Empirical Study of Reporting Practices Concerning Measurement Validity. *Educational and Psychological Measurement, 64*(4), 802-812.

Hogan, T. P. (2003). *Psychological testing: A practical introduction.* New York: Wiley.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods; Frequency of use of various types. *Educational and Psychological Measurement, 60,* 523-531,

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

Horn, J. L. (1998). A basis for research on age differences in cognitive capabilities. In J. J. McArdle & R. W. Woodcock (Eds.). *Human cognitive abilities in theory and practice* (pp.57-91). Chicago, IL: Riverside.

Hom, J. (2003). Forensic neuropsychology: Are we there yet? *Archives of Clinical Neuropsychology*, *18*, 827–845.

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

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn & Bacon.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliff NJ: Prentice Hall.

Horn, H., & Noll, J. (1994). A system for understanding cognitive capacities: A theory and the evidence on which it is based. In D. K. Detterman (Ed.).

Current topics in human intelligence Volume 4: Theories of Intelligence (pp.151-203). Norwood, NJ: Ablex.

Howell, D. C. (1997). *Statistical Methods for Psychology* (4th ed.). Melbourne: Duxbury Press.

Iezzoni, L. I., McCarthy, E. P., Davis, R. B., & Siebens, H. (2001). Mobility difficulties are not only a problem of old age. *Journal of General Internal Medicine*, *16*, 235–243.

Ingraham, L. J., & Aikken, C. B. (1996). An empirical-approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology, 10,* 120-124.

Jacobson, M. W., Delis, D. C., Hamilton, J. M., Bondi, M. W., & Salmon, D. P. (2004). How do neuropsychologists define cognitive constructs? Further thoughts on limitations of factor analysis used with normal or mixed clinical populations. *Journal of International Neuropsychological Society, 10,* 1020-1021.

Jacoby, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behaviour, 22,* 485-508.

Jones, J. J., van Shaik, P., & Witts, P. (2006). A factor analysis of the Wechsler Adult Intelligence Scale, 3rd Edition (WAIS-III) in a low IQ sample. *British Journal of Clinical Psychology, 45*(2), 145-152.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Texas: Holt Rinehart, and Winston.

Kalachstein, A. D., van Gorp, W. G., & Rapport, L. J. (1998). Variability in clinical classification of raw test scores across normative data sets. *The Clinical Neuropsychologist, 12*(3), 339-347.

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin, 112*(3)*,* 527-535.

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, R. L. (1991). Standardized and Flexible Batteries in Neuropsychology: An Assessment Update. *Neuropsychology Review, 2*(4), 281-339.

Kaneko, T., Momose, M., & Kadoya, M. (2005). Neuroimaging tools to rate cognitive impairment *Psychogeriatrics 5*, 89-92.

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications, and issues* (6th ed.). Belmont, CA: Wadsworth/Thomson.

Kaufman, A. S., & Lichtenberger, E. O. (2002). *Assessing Adolescent and Adult Intelligence* (2nd ed.). London: Allyn and Bacon.

Kelly, T. L. (1927). *Interpretation of educational measurements.* New York: Tarrytown-on Hudson.

Kennepohl, S., Shore, D., Nabors, N., & Hanks, R. (2004). African American acculturation and neuropsychological test performance following traumatic brain injury. *Journal of the International Neuropsychological Society*, *10*, 566–577.

Kieffer, K. M. (1999). An introductory primer on the appropriate use of exploratory and confirmatory factor analysis. *Research in the Schools, 6,* 75-92.

Kline, P. (2000). *Handbook of psychological testing.* New York: Routledge.

Kolb, B., & Whishaw, I. Q. (1996). *Fundamentals of Human Neuropsychology* (4th ed.). New York: W.H. Freeman and Company.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Larrabee, G. L. (2003). Lessons on measuring construct validity: A commentary on Delis, Jacobson, Bondi, Hamilton and Salmon. *Journal of the International Neuropsychological Society, 9,* 947-953.

Leaper, S. A., Murray, A. D., Lemmon, H. A., Staff, R. T., Deary, I. J., Crawford, J. R., et al. (2001). Neuropsychological Correlates of Brain White Matter Lesions Depicted on MR Images: 1921 Aberdeen Birth Cohort. *Neuroradiology, 221*, 51-55.

Lees-Haley, P. R., Smith, H. H., Williams, C. W., & Dunn, J. T. (1995). Forensic Neuropsychological Test Usage: An Empirical Study. *Archives of Clinical Neuropsychology, 11*(1), 45-51.

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

Lezak, M. D. (1983). *Neuropsychological Assessment* (2nd ed.). Oxford: Oxford University Press.

Lezak, M. D. (2002). Responsive Assessment and the Freedom to Think for Ourselves. *Rehabilitation Psychology, 47*(3), 339-353.

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

Linn, R. I., & Gronlund, N. E. (1995). *Measurement in assessment and teaching* (7th ed.). Englewood Cliffs, NJ: Merrill.

Liotti, M., & Mayberg, H. S. (2001). The Role of Functional Neuroimaging in the Neuropsychology of Depression. *Journal of Clinical and Experimental Neuropsychology, 23*(1), 121-136.

Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika, 20,* 1-22.

Lord, F. M. (1959). An approach to mental test theory. *Psychometrika, 24,* 283-303.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Lucke, J. F. (2005). The a and the w of Congeneric Test Theory: An Extension of Reliability and Internal Consistency to Heterogeneous Tests. *Applied Psychological Measurement, 29*(1), 65-81.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*, 161-175.

Manly, J. J., & Echemendia, R. J. (2007). Race-specific norms: Using the model of hypertension to understand issues of race, culture, and education in neuropsychology. *Archives of Clinical Neuropsychology, 22,* 319-325.

Manly, J. J. (2005). Advantages and disadvantages of separate norms for African Americans. *The Clinical Neuropsychology, 19*(2), 270-275.

Manly, J. J., Jacobs, D. M., Touradji, P., Small, S. A. & Stern, Y. (2002). Reading level attenuates differences in neuropsychological test performance between African American and White elders. *Journal of the International Neuropsychological Society*, *8*(3), 341–348.

Manly, J. J., Touradji, P., Tang, M.-X., & Stern, Y. (2003). Literacy and memory decline among ethnically diverse elders. *Journal of Clinical and Experimental Neuropsychology. Special Issue: Cognitive Reserve*, *25*(5), 680–690.

Matarazzo, J. D. (1990). Psychological Assessment Versus Psychological Testing Validation From Binet to the School, Clinic, and Courtroom. *American Psychologist, 45*(9), 999-1017.

McCarthy, R. A., & Warrington, E. K. (1990). *Cognitive Neuropsychology.* San Diego: Academic Press.

McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires* (2nd ed.). New York: Oxford University Press.

McKenna, P., & Warrington, E. K. (1996). The Analytical Approach to Neuropsychological Assessment. In I. Grant & K. M. Adams (Eds.). *Neuropsychological Assessment of Neuropsychiatric Disorders* (2nd ed.). New York: Oxford University Press.

Michell, J. (1986). Measurement Scales and Statistics: A Clash of Paradigms *Psychological Bulletin, 100*(3), 398-407.

Microsoft Office Excel 2003 [Computer Software]. (2003). Sydney, Australia: Microsoft Pty. Limited.

Milberg, W. P., Hebben, N., & Kaplan, E. (1996). The Boston Process Approach to Neuropsychological Assessment. In I. Grant & K. M. Adams (Eds.). *Neuropsychological Assessment of Neuropsychiatric Disorders* New York: Oxford University Press.

Miller, L. S., & Rohling, M. L. (2001). A Statistical Interpretive Method for Neuropsychological Test Data. *Neuropsychology Review, 11*, 143-168.

Millis, S. R., Malina, A. C., Bowers, D. A., & Ricker, J. H. (1999). Confirmatory factor analysis of the Wechsler Memory Scale - III. *Journal of Clinical and Experimental Neuropsychology, 21*(1), 87-93.

Mitrushina, M., Boone, K. B., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). Oxford: Oxford University Press.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika, 8,* 161-168.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioural Research, 22,* 267-305.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological Testing: Principles and Applications* (6th ed.). Sydney: Pearson Education International.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures.* Washington, DC:American Psychological Association.

Nichols, P., & Kuehl, B. J. (1999). Prophesying the Reliability of Cognitively Complex Assessments. *Applied Measurement in Education 12*(1), 73-94.

Nirini, R., Caramelli, P., Herrera, E., Porto, C. S., Charchat-Fichman, H., Carthery, M. T., et al. (2004). Performance of illiterate and literate nondemented elderly subjects in two tests of long-term memory. *Journal of the International Neuropsychological Society*, *10*, 634–638.

Novick, M. R. (1966). The Axioms and Principal Results of Classical Test Theory. *Journal of Mathematical Psychology, 3,* 1-18.

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

Nunnally, J. C. (1978). *Psychometric Theory*. Sydney: McGraw-Hill Book Company.

Ogden, J. A. (1996). Fractured Minds: A case-study approach to clinical neuropsychology. Oxford: Oxford University Press.

Olm, T. K., & Senior, G. J. (2006). A Reliable Approach to Psychological Assessment (RAPT). Workshop for "Forensic Neuropsychology: Foundations and Practice" the 12th Annual Conference of the APS College of Clinical Neuropsychologists.

Osburn, H. G. (2000). Coefficient Alpha and Related Internal Consistency Reliability Coefficient. *Psychological Methods, 5*(3), 343-355.

Ostrosky-Sol´ıs, F., Ramirez, M., & Ardila, A. (2004). Effects of culture and education on neuropsychological testing: A preliminary study with

indigenous and nonindigenous population. *Applied Neuropsychology*, *11*(4), 186–193.

Park, L. T., & Gonzalez, R. G. (2004). Computed Tomography. In D. D. Dougherty, S. L. Rauch & J. G. Rosenbaum (Eds.). *Essentials of Neuroimaging for Clinical Practice* (pp. 1-20). Arlington, VA American Psychiatric Publishing.

Park, W. L., Mayer, R. S., Moghimi, C., Park, J. M., & Deremeik, J. T. (2005). Rehabilitation of hospital inpatients with visual impairments and disabilities from systemic illness. *Archives of Physical Medicine and Rehabilitation*, *86*, 79–81.

Payne, R. W., & Jones, H. G. (1957). Statistics for the Investigation of Individual Cases. *Journal of Clinical Psychology, 18*, 115-121.

Price, L. R., Tulsky, D. S., Millis, S. R., & Weiss, L. (2002). Redefining the factor structure of the Wechsler Memory Scale - III: Confirmatory factor analysis with cross-validation. *Journal of Clinical and Experimental Neuropsychology, 24,* 574-585.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*, 33-65.

Raykov, T. (1998). Coefficient Alpha and Composite Reliability with Interrelated Nonhomogenous Items. *Applied Psychological Measurement, 22*(4), 375-385.

Reed, J. E. (1996). Fixed vs. Flexible Neuropsychological Test Batteries Under the Daubert Standard for the Admissibility of Scientific Evidence. *Behavioural Sciences and the Law, 14*, 315-322.

Reed, J. E. (1999). Current Status of the Admissibility of Expert Testimony After *Daubert* and *Joiner Journal of Forensic Neuropsychology, 1*(1), 49-67.

Reinhardt, B. (1996). Factors Affecting Coefficient Alpha: A Mini Monte Carlo Study. *Advances in Social Science Methodology, 4,* 3-20.

Reis, A., Guerreiro, M.,&Petersson, K. M. (2003). A sociodemographic and neuropsychological characterization of an illiterate population. *Applied Neuropsychology*, *10*(4), 191–204.

Reitan, R. M. (1986). *The Neuropsychological Deficit Scale for adults: Computer program.* Tucson, AZ: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1988). *Traumatic Brain Injury* (Vol 2.). Tucson, AZ: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

Reitan, R. M. (1955). Certain differential effects of left and right cerebral lesions in human adults. *Journal of Comparative and Physiological Psychology, 48,* 474-477.

Reitan, R. M., & Wolfson, D. (1996). Theoretical, Methodological and Validational Bases of the Halstead-Reitan Neuropsychological Test Battery. In I. Grant & K. M. Adams (Eds.), *Neuropsychological Assessment of Neuropsychiatric Disorders* (2nd ed., pp. 3 - 42). New York Oxford University Press.

Reitan, R. M., & Wolfson, D. (2001). Critical evaluation of "Assessment: neuropsychological testing of adults". *Archives of Clinical Neuropsychology, 16*, 215-226.

Reitan, R. M., & Wolfson, D. (2004). The Halstead-Reitan Neuropsychological Test Battery for Adults: Theoretical, Methodological and Validational Bases. In G.

Goldstein & S. Beers (Eds.). *Comprehensive Handbook of Psychological Assessment. Volume 1. Intellectual and Neuropsychological Assessment* (Vol. 1). Hoboken, NJ: John Wiley and Sons.

Reitan, R.M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation.* Tucson, AZ: Neuropsychology Press.

Rojas, D. C., & Bennett, T. L. (1995). Single Versus Composite Score Disciminative Validity With the Halstead-Reitan Battery and the Stroop Test in Mild Brain Injury. *Archives of Clinical Neuropsychology, 10*(2), 101-110.

Rohling, M. L., & Miller, L. S. (2004). *Rohling's Interpretive Method: Use of Meta-Analytic Procedures for Single Case Data Analysis.* Workshop presented at the 24th Annual Conference of the National Academy of Neuropsychology (NAN), Seattle, November, 2004.

Rohling, M. L., Miller, L. S., & Langhinrichsen-Rohling, J. (2004). Rohling's Interpretive Method for Neuropsychological Case Data: A Response to Critics. *Neuropsychology Review, 14*(3), 155-169.

Rohling, M. L., Langhinrichsen-Rohling, J., & Miller, L. S. (2003). Actuarial assessment of malingering: The RIM Process. In R. D. Franklin (Ed.). *Prediction in Forensic and Neuropsychology: Sound Statistical Procedures,* Mahwah, NJ: Erlbaum.

Rohling, M. L., Williamson, D. J., Miller, L. S., & Adams, R. (2003). Using the Halstead Reitan Battery to diagnose brain damage: A comparison of the predictive power of traditional techniques to Rohling's Interpretive Method. *The Clinical Neuropsychologist, 17,* 531-544.

Rosenthal, B., Sands, S. A., & Van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology, 15*(4), 349-359.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioural research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Ruscio, J., & Ruscio, A. M. (2002). A Structure-Based Approach to Psychological Assessment: Matching Measurement Models to Latent Structure. *Assessment, 9*(1), 40-16.

Russell, E. W. (1995). The accuracy of automated and clinical detection of brain damage and lateralization in neuropsychology. *Neuropsychology Review, 5*(1), 1-68.

Russell, E. W. (1997). Developments in the psychometric foundations of neuropsychological assessment. In G. Goldstein & T. M. Incagnoli (Eds.). *Contemporary approaches to neuropsychological assessment* (pp.15-65). New York: Plennum.

Russell, E. W. (1998). In defence of the Halstead Reitan Battery: A critique of Lezak's review. *Archives of Clinical Neuropsychology, 13,* 353-381.

Russell, E. W. (2000). The Cognitive-Metric, Fixed Battery Approach to Neuropsychological Assessment In R. D. Vanderploeg (Ed.). *Clinician's Guide to Neuropsychological Assessment* (2nd. ed., pp. 449-481). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Russell, E. W. (2007). Commentary on "A motion to exclude and the 'fixed' versus 'flexible' debate in 'forensic' neuropsychology. *Archives of Clinical Neuropsychology, 22,* 787-790.

Russell, E. W., & Russell, S. L. K. (2003). Twenty ways and more of diagnosing brain damage where there is none. (errors sometimes committed by neuropsychologists). *Journal of Controversial Medical Claims, 10*(1), 1-14.

Russell, E. W., Russell, S. L., & Hill, B. D. (2005). The fundamental psychometric status of neuropsychological batteries. *Archives of Clinical Neuropsychology, 20,* 785-794.

Russell, E. W., & Starkey, R. I. (2001). *Halstead, Russell neuropsychological evaluation system – revised [manual and computer program]*. Los Angeles: Western Psychological Services.

Ryan, J. J., & Paola, A. M. (2001). Exploratory factor analysis of the WAIS-III in a mixed clinical sample. *Archives of Clinical Neuropsychology, 16*(2), 151-156.

Ryan, E. L., Baird, R., Mindt, M. R., Byrd, D., Monzones, J., & Morgellow, S. (2005). Neuropsychological impairment in racial/ethnic minorities with HIV infection and low literacy levels: Effects of education and reading level in participant characterization. *Journal of the International Neuropsychological Society*, *11*, 889–898.

Saklofske, D.H., Hildebrand, D. K., & Gorsuch, R. L. (2000). Replication of the factor structure of the Wechsler Adult Intelligence Scale – Third Edition with a Canadian sample. *Psychological Assessment, 12*(4), 436-439.

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

Sattler, J. M., & Ryan, J. J. (1998). *Assessment of Children, Revised and Updated Third Edition WAIS-III Supplement.* San Diego: Jerome M. Sattler, Publisher.

Savoy, R. L., & Gollub, R. L. (2004). Functional Magnetic Resonance Imaging In D. D. Dougherty, S. L. Rauch & J. G. Rosenbaum (Eds.). *Essentials of*

Neuroimaging for Clinical Practice (pp. 93-104). Arlington, VA: American Psychiatric Publishing.

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Educational and Psychological Measurement, 60*, 157-173.

Schmitt, N. (1996). Uses and Abuses of Coefficient Alpha. *Psychological Assessment, 8*(4), 350-353.

Schretlen, D., Benedict, R. H. B., & Bobholz, J. H. (1994). Composite Reliability and Standard Errors of Measurement for a Seven-subtest Short Form of the Wechsler Adult Intelligence Scale – Revised. *Psychological Assessment, 6*(3), 188-190.

Seretny, M. L., Dean, R. S., Gray, J. W., & Hartlage, L. C. (1986). The Practice of Clinical Neuropsychology in the United States. *Archives of Clinical Neuropsychology, 1*, 5-12.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer.* Newbury Park CA: Sage.

Silverstein, A. B. (1981). Reliability and Abnormality of Test Score Differences. *Journal of Clinical Psychology*, *37*(2), 392-394.

Smith, A. (1982). *Symbol Digit Modalities Test, Revised, Manual.* Los Angeles: Western Psychological Services.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry: the principles and practices of statistics in biological research* (3rd ed.). New York: W. H. Freeman and Co.

Spearman, C. (1904). "General Intelligence" Objectively Determined and Measured. *American Journal of Psychology, 15*, 201-293.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: administration, norms, and commentary* (2nd ed.). . New York: Oxford University Press.

Stankov, L., & Horn, J. L. (1980). Human abilities revealed through auditory tests. *Journal of Educational Psychology, 72,* 19-42.

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

Sternberg, R. J. (Ed.) (1994). *Encyclopedia of human intelligence* (Vol.2). New York: Macmillan.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms and Commentary* (3rd ed.). New York: Oxford University Press.

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment, 80*(1), 99-103.

Sullivan, K., & Bowden, S. C. (1997). Which Tests Do Neuropsychologists Use? *Journal of Clinical Psychology, 53*(7), 657-661.

Sweet, J. J., Moberg, P. J., & Suchy, Y. (2000). Ten-Year Follow-up Survey of Clinical Neuropsychologists: Part I. Practices and Beliefs. *The Clinical Neuropsychologist, 14*(1), 18-37.

Sweet, J. J., Nelson, N. W., & Moberg, P. J. (2006). The TCN/AACN 2005 "Salary Survey": Professional Practices, Beliefs and Incomes of U.S. Neuropsychologists. *The Clinical Neuropsychologist 20*, 325-354.

Sweet, J., & Moberg, P. (1990). A survey of practices and beliefs among ABPP and non-ABPP clinical neuropsychologists. *The Clinical Neuropsychologist, 4*, 101-120.

Sweet, J., Moberg, P., & Westergaard, C. (1996). Five-year follow-up survey of practices and beliefs of clinical neuropsychologists. *The Clinical Neuropsychologist, 10*, 201-221.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). Sydney: Allyn and Bacon.

Tellegen, A., & Briggs, P. F. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting and Clinical Psychology, 31,* 499-506.

Thompson, B. (1991). Review of generalizability theory: A primer by R. J. Shavelson & N. W. Webb. *Educational and Psychological Measurement, 51,* 1069-1075.

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

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is Datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174-195.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educaitonal and Psychological Measurement, 56,* 197-208.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.). *Educational measurement* (pp.560-620). Washington, DC: American Council on Education.

Tranel, D. (1996). The Iowa-Benton School of Neuropsychological Assessment. In I. Grant & K. M. Adams (Eds.). *Neuropsychological Assessment of Neuropsychiatric Disorders* (2nd ed., pp. 81-101). New York: Oxford University Press.

Traub, R. E., & Rowley, G. L. (1991). An NCME Instructional Module on Understanding Reliability. *Educational Measurement: Issues and Practice 10*(1), 37-44.

Tulsky, D. S., & Price, L. R. (2003). The Joint WAIS-III and WMS-III Factor Structure: Development and Cross-Validation of a Six-Factor Model of Cognitive Functioning. *Psychological Assessment, 15*(2), 149-162.

Tulsky, D. S., Chiaravalloti, N. D., Palmer, B. W., & Chelune, G. J. (2003). The Wechsler Memory Scale, Third Edition: A New Perspective (pp. 95-133). In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik, R. Bornstein, A. Prifitera, M. F. Ledbetter (Eds.). *Clinical Interpretation of the WAIS-III and WMS-III*. Sydney: Academic Press.

Tulsky, D. S., Ivnik, R., J , Price, L. R., & Wilkins, C. (2003). Assessment of Cognitive Functioning with the WAIS-III and WMS-III: Development of a Six-Factor Model (p.147-179). In D. S. Tulsky, D. H. Saklofske, R. K. Heaton, R. Bornstein, M. F. Ledbetter, G. J. Chelune, R. J. Ivnik & A. Prifitera (Eds.). *Clinical Interpretation of the WAIS-III and WMS-III*. Sydney: Academic Press.

Tulsky, D. S., Rolfhus, E. L., & Zhu, J. (2000). Two-Tailed Versus One-Tailed Base Rates of Discrepancy Scores on the WAIS-III. *The Clinical Neuropsychologist, 14*(4), 451-460.

Tulsky, D. S., Zhu, J., & Vasquez, C. (1998). The clinical utility of the WAIS-III IQ and index scores in patients with neuropsychological disorders. Abstract

presentation at the 26th Annual Convention of the International Neuropsychological Society. Honolulu. *Journal of the International Neuropsychological Society, 3*.

Vacha-Haase, T. Kogan, L. R., & Thompson, B. (2000). Sample Compositions and Variables in Published Studies Versus Those in Test Manuals: Validity of Score Reliability Inductions. *Educational and Psychological Measurement, 60*(4), 509-522.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability Generalization: Moving Towards Improved Understanding and Use of Score Reliability. *Educational and Psychological Measurement, 62*(4), 562-569.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices Regarding Reporting of Reliability Coefficients: A Review of Three Journals. *The Journal of Experimental Education, 67*(4), 335-341.

Vanderploeg, R. D. (2000). Interviewing and Testing: The Data Collection Phase of Neuropsychological Evaluations. In R. D. Vanderploeg (Ed.). *Clinician's Guide to Neuropsychological Assessment* (2nd ed., pp. 3-38). Mahway, New Jersey Lawrence Erlbaum Associates.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*(1), 1-28.

Wechsler, D. (1939). *The Measurement of Adult Intelligence*. Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised manual*. New York: Psychological Corporation.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale-Third Edition and Wechsler Memory Scale, Third Edition Technical Manual.* San Antonia, Texas: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Adult Intelligence Scale – Third Edition Administration and Scoring Manual.* San Antonio, Texas: The Psychological Corporation.

Wechsler, D. (1997c). *Wechsler Memory Scale – Third Edition Administration and Scoring Manual.* San Antonio, Texas: The Psychological Corporation.

Wedding, D., & Faust, D. (1989). Clinical Judgment and Decision Making in Neuropsychology. *Archives of Clinical Neuropsychology, 4*, 233-265.

Weiner, E. A., & Stewart, B. J. (1984). *Assessing individuals: Psychological and educational tests and measurement.* Boston: Little, Brown.

Whittle, C., Corrada, M. M., Dick, M., Ziegler, R., Kahle-Wrobleski, K., Paganini-Hill, A., & Kawas, C. (2007). Neuropsychological data in nondemented oldest old: The 90+ Study. *Journal of Clinical and Experimental Neuropsychology, 296*(3), 290-299.

Wilde, N. J., Strauss, E., Chelune, G. J., Hermann, B. P., Hunter, M., Loring, D. W., Martin, R. C., & Sherman, E. M. S. (2003). Confirmatory factor analysis of the WMS-III in patients with temporal lobe epilepsy. *Psychological Assessment, 15,* 56-63.

Wilkinson, G. S. (1993). *Wide Range Achievement Test 3*. Wilmington, DE: Wide Range Inc.

Wilkinson, G. S., & Robertson, G. J. (2006). *WRAT-4 Wide Range Achievement Test: Professional Manual.* Florida: Psychological Assessment Resources.

Williams, J. M. (1997). The prediction of premorbid memory ability. *Archives of Clinical Neuropsychology, 12,* 745-756.

Williams, V. G., Bruce, J. M., Westervelt, H. J., David, J. D., Grace, J., Malloy, P. F., & Tremont, G. (2007). Boston naming performance distinguishes between Lewy body and Alzheimer's dementias. *Archives of Clinical Neuropsychology, 22,* 925-931.

Wilson, N. (1998). Educational Standards and the Problem of Error. *Education Policy Analysis Archives, 6*(10), Retrieved 20th December, 2006 from http://epaa.asu.edu/epaa/v6n10/.

Woodcock, R. W., & Mather, N. (1989). *Woodcock-Johnson Tests of Achievement.* Allen, TX: DLM Teaching Resources.

Yancey, S. W., & Phelps, E. A. (2001). Functional Neuroimaging and Episodic Memory: A Perspective. *Journal of Clinical and Experimental Neuropsychology, 23*(1), 32-48.

Zwick, W. R., & Velicer, W. F. (1986). A comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

APPENDIX A:

Verbal Fluency Composite Calculations

Subtest Observed Score (Formula 5.2):

$$X_{VFC} = 5 + 6$$
$$X_{VFC} = 11$$

Composite Mean (Formula 5.3a):

$$M_{VFC} = 10(2)$$
$$M_{VFC} = 20$$

Composite Standard Deviation (Formula 5.5):

$$SD_{VFC} = \sqrt{3^2 + 3^2 + 2(3)(3)(0.402)}$$
$$SD_{VFC} = \sqrt{18 + 7.236}$$
$$SD_{VFC} = \sqrt{25.236}$$
$$SD_{VFC} = 5.0245$$

Composite Deviation Quotient (Formula 5.8a):

$$DQ_{VFC} = (11 - 20/5.024) \times 15 + 100$$
$$DQ_{VFC} = 73.1290 = 73$$

Composite Reliability Coefficient (Formula 5.6):

$$r_{VFC} = \frac{(0.82 + 0.55) + [(2) * (0.402)]}{2 + [(2) * (0.402)]}$$
$$r_{VFC} = \frac{1.37 + 0.804}{2 + 0.804}$$
$$r_{VFC} = \frac{2.174}{2.804}$$
$$r_{cc} = 0.7753 = 0.775$$

Standard Error of the Estimate (Formula 5.9a):

$$SE_{E_{VFC}} = 15\sqrt{(0.775)(1 - 0.775)}$$
$$SE_{E_{VFC}} = 15\sqrt{0.1744}$$
$$SE_{E_{VFC}} = 6.2637 = 6.264$$

Predicted True Score (Formula 5.11a):

$$DQ_{PT_{VFC}} = 0.775(73.129 - 100) + 100)$$
$$DQ_{PT_{VFC}} = 79.1749$$

Ninety Percent Confidence Interval (Formula 5.10):

$$79.1749 \pm (1.64) * (6.264)$$
69 - 89

Standard Error of Prediction (Formula 5.12a):

$$SE_{PVFC} = 15\sqrt{(1 - 0.775^2)}$$
$$SE_{PVFC} = 15\sqrt{0.3994}$$
$$SE_{PVFC} = 9.4794 = 9.480$$

Ninety Percent Re-test Confidence Interval (Formula 5.10a):

$$79.1749 \pm (1.64) * (9.480)$$
64 - 95

APPENDIX B:

Reliability Simulation Calculations

Intercorrelations are held constant at 0.5 for all subtests and computations make use of formula 5.6.

Composite 1:
Two subtest composite with component subtest reliability coefficients r = .5

$$\begin{aligned} r_c &= [(0.5 + 0.5) + (2)(0.5)]/[2 + (2)(0.5)] \\ &= [1 + 1]/[2 + 1] \\ &= 2 / 3 \\ &= 0.67 \end{aligned}$$

Composite 2:
Two subtest composite with component subtest reliability coefficients r = .6

$$\begin{aligned} r_c &= [(0.6 + 0.6) + (2)(0.5)]/[2 + (2)(0.5)] \\ &= [1.2 + 1]/[2 + 1] \\ &= 2.2 / 3 \\ &= 0.73 \end{aligned}$$

Composite 3:
Two subtest composite with component subtest reliability coefficients r = .7

$$\begin{aligned} r_c &= [(0.7 + 0.7) + (2)(0.5)]/[2 + (2)(0.5)] \\ &= [1.4 + 1]/[2 + 1] \\ &= 2.4 / 3 \\ &= 0.8 \end{aligned}$$

Composite 4:
Two subtest composite with component subtest reliability coefficients r = .8

$$\begin{aligned} r_c &= [(0.8 + 0.8) + (2)(0.5)]/[2 + (2)(0.5)] \\ &= [1.6 + 1]/[2 + 1] \\ &= 2.6 / 3 \\ &= 0.87 \end{aligned}$$

Composite 5:
Two subtest composite with component subtest reliability coefficients r = .9

$$\begin{aligned} r_c &= [(0.9 + 0.9) + (2)(0.5)]/[2 + (2)(0.5)] \\ &= [1.8 + 1]/[2 + 1] \\ &= 2.8 / 3 \\ &= 0.93 \end{aligned}$$

APPENDIX C:

Simulated Composite Reliability for two, three, four, five, six, seven, eight, nine and ten subtest composites with intercorrelations held constant at .5

Table C. 1 Simulated Composite Reliability for Two, Three, Four, Five, Six, Seven, Eight, Nine and Ten Subtest Composites with Intercorrelations Held Constant at .5

	Composite Group 1 α $r_{kk'} = .5$ $r_{k(k-1)} = .5$	**Composite Group 2 α** $r_{kk'} = .6$ $r_{k(k-1)} = .5$	**Composite Group 3 α** $r_{kk'} = .7$ $r_{k(k-1)} = .5$	**Composite Group 4 α** $r_{kk'} = .8$ $r_{k(k-1)} = .5$	**Composite Group 5 α** $r_{kk'} = .9$ $r_{k(k-1)} = .5$
2 Subtest Composite	.67	.73	.80	.87	.93
3 Subtest Composite	.75	.80	.85	.90	.95
4 Subtest Composite	.80	.84	.88	.92	.96
5 Subtest Composite	.83	.87	.91	.93	.97
6 Subtest Composite	.86	.89	.91	.94	.97
7 Subtest Composite	.88	.90	.93	.95	.98
8 Subtest Composite	.89	.91	.93	.96	.98
9 Subtest Composite	.90	.92	.94	.96	.98
10 Subtest Composite	.91	.93	.95	.96	.98

Composite Group 1: Two to ten subtest composites with intercorrelations between component subtests held at r = .5 and subtest reliability coefficients of r = .5.

Composite Group 2: Two to ten subtest composites with intercorrelations between component subtests held at r = .5 and subtest reliability coefficients of r = .6.

Composite Group 3: Two to ten subtest composites with intercorrelations between component subtests held at r = .5 and subtest reliability coefficients of r = .7.

Composite Group 4: Two to ten subtest composites with intercorrelations between component subtests held at r = .5 and subtest reliability coefficients of r = .8.

Composite Group 5: Two to ten subtest composites with intercorrelations between component subtests held at r = .5 and subtest reliability coefficients of r = .9.

APPENDIX D:

Simulated Composite Reliability for two, three, four, five, six, seven, eight, nine and ten subtest composites with intercorrelations held constant at .7

Table D. 1Simulated Composite Reliability for Two, Three, Four, Five, Six, Seven, Eight, Nine and Ten Subtest Composites with Intercorrelations Held Constant at .7

	Composite Group 1 α $r_{kk'} = .5$ $r_{k(k-1)} = .7$	**Composite Group 2 α** $r_{kk'} = .6$ $r_{k(k-1)} = .7$	**Composite Group 3 α** $r_{kk'} = .7$ $r_{k(k-1)} = .7$	**Composite Group 4 α** $r_{kk'} = .8$ $r_{k(k-1)} = .7$	**Composite Group 5 α** $r_{kk'} = .9$ $r_{k(k-1)} = .7$
2 Subtest Composite	.71	.74	.82	.88	.94
3 Subtest Composite	.79	.83	.89	.92	.96
4 Subtest Composite	.84	.87	.90	.94	.97
5 Subtest Composite	.87	.89	.92	.95	.97
6 Subtest Composite	.89	.91	.93	.96	.98
7 Subtest Composite	.90	.92	.94	.96	.98
8 Subtest Composite	.92	.93	.95	.97	.98
9 Subtest Composite	.92	.94	.95	.97	.98
10 Subtest Composite	.93	.95	.96	.97	.99

Composite Group 1: Two to ten subtest composites with intercorrelations between component subtests held at r = .7 and subtest reliability coefficients of r = .5.

Composite Group 2: Two to ten subtest composites with intercorrelations between component subtests held at r = .7 and subtest reliability coefficients of r = .6.

Composite Group 3: Two to ten subtest composites with intercorrelations between component subtests held at r = .7 and subtest reliability coefficients of r = .7.

Composite Group 4: Two to ten subtest composites with intercorrelations between component subtests held at r = .7 and subtest reliability coefficients of r = .8.

Composite Group 5: Two to ten subtest composites with intercorrelations between component subtests held at r = .7 and subtest reliability coefficients of r = .9.

APPENDIX E:

Reliability Calculations for Verbal

Composites suggested by PCA

Table E. 1 Intercorrelations between Verbal Subtests Derived from PCA sample

	BNT	WAIS3-VO	WAIS3-CO	WAIS3-SI	STW	WRAT3-Reading
WAIS3-IN	.529	.661	.554	.524	.532	.590
BNT		.524	.457	.411	.450	.457
WAIS3-VO			.659	.591	.534	.666
WAIS3-CO				.505	.425	.482
WAIS3-SI					.381	.423
STW						.611

Composite 1:
Most Valid (WAIS3-IN, BNT and WAIS3-VO)

$$r_{cc} = \frac{(0.91 + 0.81 + 0.93) + [(2) * (.529 + .661 + .524)]}{3 + [(2) * (.529 + .661 + .524)]}$$
$$r_{cc} = \frac{2.65 + [(2) * (1.714)]}{3 + [(2) * (1.714)]}$$
$$r_{cc} = \frac{6.078}{6.428}$$
$$r_{cc} = 0.9456$$

Composite 2:
Most Reliable (WRAT3-Reading, WAIS3-VO and WAIS3-IN)

$$r_{cc} = \frac{(0.95 + 0.93 + 0.91) + [(2) * (.590 + .666 + .661)]}{3 + [(2) * (.590 + .666 + .661)]}$$
$$r_{cc} = \frac{2.79 + [(2) * (1.917)]}{3 + [(2) * (1.917)]}$$
$$r_{cc} = \frac{6.624}{6.834}$$
$$r_{cc} = 0.9693$$

Composite 3:
Most Practical (WRAT3-Reading, STW and BNT)

$$r_{cc} = \frac{(0.95 + 0.83 + 0.81) + [(2) * (.450 + .457 + .611)]}{3 + [(2) * (.450 + .457 + .611)]}$$
$$r_{cc} = \frac{2.59 + [(2) * (1.518)]}{3 + [(2) * (1.518)]}$$
$$r_{cc} = \frac{5.626}{6.036}$$
$$r_{cc} = 0.9321$$

APPENDIX F

Comparison of WRAT-4 Reading Composite with RAPT Reading Composite.

Table F. 1WRAT-4 Reading Composites

	RAPT Reading Composite			**WRAT-4 Reading Composite**			
		90% Confidence Interval			*90% Confidence Interval*		
Raw Score	DQ	Low	High	Standard Score	Low	High	**Discrepancy**
110	51	48	58	55	51	61	-4
111	51	49	59	56	52	62	-5
112	52	49	59	56	52	62	-4
113	53	50	60	57	53	63	-4
114	53	50	60	57	53	63	-4
115	54	51	61	57	53	63	-3
116	54	51	61	58	54	65	-4
117	55	52	62	58	54	65	-3
118	55	52	62	59	55	65	-4
119	56	53	63	59	55	65	-3
120	56	53	63	59	55	65	-3
121	57	54	64	60	56	66	-3
122	57	54	64	60	56	66	-3
123	58	55	65	61	57	67	-3
124	59	55	66	61	57	67	-2
125	59	56	66	62	58	68	-3
126	60	56	67	62	58	68	-2
127	60	57	67	62	58	68	-2
128	61	57	68	63	59	69	-2
129	61	58	68	63	59	69	-2
130	62	58	69	64	60	70	-2
131	62	59	69	64	60	70	-2
132	63	59	70	65	61	71	-2
133	63	60	70	65	61	71	-2
134	64	60	71	65	61	71	-1
135	65	61	71	66	62	72	-1
136	65	62	72	66	62	72	-1
137	66	62	72	67	63	73	-1
138	66	63	73	67	63	73	-1
139	67	63	73	68	64	74	-1
140	67	64	74	68	64	74	-1
141	68	64	74	69	65	75	-1
142	68	65	75	69	65	75	-1
143	69	65	75	70	66	76	-1
144	69	66	76	70	66	76	-1
145	70	66	76	70	66	76	0
146	71	67	77	71	67	77	0
147	71	67	77	71	67	77	0
148	72	68	78	72	68	78	0
149	72	68	79	72	68	78	0
150	73	69	79	73	69	78	0
151	73	69	80	73	69	78	0
152	74	70	80	74	70	79	0
153	74	70	81	74	70	79	0
154	75	71	81	75	71	80	0
155	75	71	82	75	71	80	0
156	76	72	82	76	72	81	0
157	77	72	83	76	72	81	1

Table F.1 (continued)

	RAPT Reading Composite			WRAT-4 Reading Composite			
		90% Confidence Interval			*90% Confidence Interval*		
Raw Score	DQ	Low	High	Standard Score	Low	High	**Discrepancy**
158	77	73	83	77	73	82	0
159	78	74	84	77	73	82	1
160	78	74	84	78	74	83	0
161	79	75	85	78	74	83	1
162	79	75	85	79	74	84	0
163	80	76	86	79	74	84	1
164	80	76	86	80	75	85	0
165	81	77	87	80	75	85	1
166	81	77	87	81	76	86	0
167	82	78	88	81	76	86	1
168	83	78	88	82	77	87	1
169	83	79	89	82	77	87	1
170	84	79	89	83	78	88	1
171	84	80	90	83	78	88	1
172	85	80	91	84	79	89	1
173	85	81	91	84	79	89	1
174	86	81	92	85	80	90	1
175	86	82	92	86	81	91	0
176	87	82	93	86	81	91	1
177	87	83	93	87	82	92	0
178	88	83	94	87	82	92	1
179	89	84	94	88	83	93	1
180	89	84	95	88	83	93	1
181	90	85	95	89	84	94	1
182	90	86	96	89	84	94	1
183	91	86	96	90	85	95	1
184	91	87	97	90	85	95	1
185	92	87	97	91	86	96	1
186	92	88	98	91	86	96	1
187	93	88	98	92	87	97	1
188	93	89	99	93	88	98	0
189	94	89	99	94	89	99	0
190	95	90	100	94	89	99	1
191	95	90	100	94	89	99	1
192	96	91	101	95	90	100	1
193	96	91	101	95	90	100	1
194	97	92	102	96	91	101	1
195	97	92	103	96	91	101	1
196	98	93	103	97	92	102	1
197	98	93	104	98	93	103	0
198	99	94	104	99	94	104	0
199	99	94	105	99	94	104	0
200	100	95	105	100	95	105	0
201	101	95	106	100	95	105	1
202	101	96	106	100	95	105	1

Table F.1 (continued)

Raw Score	RAPT Reading Composite			WRAT-4 Reading Composite			Discrepancy
		90% Confidence Interval			*90% Confidence Interval*		
	DQ	Low	High	Standard Score	Low	High	
203	102	96	107	101	96	106	1
204	102	97	107	101	96	106	1
205	103	97	108	102	97	107	1
206	103	98	108	103	98	108	0
207	104	99	109	103	98	108	1
208	104	99	109	104	99	109	0
209	105	100	110	104	99	109	1
210	105	100	110	105	100	110	0
211	106	101	111	105	100	110	1
212	107	101	111	106	101	111	1
213	107	102	112	106	101	111	1
214	108	102	112	107	102	112	1
215	108	103	113	108	103	113	0
216	109	103	113	108	103	113	1
217	109	104	114	109	104	114	0
218	110	104	114	109	104	114	1
219	110	105	115	110	105	115	0
220	111	105	116	110	105	115	1
221	111	106	116	111	106	116	0
222	112	106	117	111	106	116	1
223	113	107	117	112	107	117	1
224	113	107	118	113	108	118	0
225	114	108	118	113	108	118	1
226	114	108	119	114	109	119	0
227	115	109	119	114	109	119	1
228	115	109	120	115	110	120	0
229	116	110	120	115	110	120	1
230	116	111	121	116	111	121	0
231	117	111	121	116	111	121	1
232	117	112	122	117	112	122	0
233	118	112	122	117	112	122	1
234	119	113	123	118	113	123	1
235	119	113	123	119	114	124	0
236	120	114	124	119	114	124	1
237	120	114	124	120	115	125	0
238	121	115	125	120	115	125	1
239	121	115	125	121	116	126	0
240	122	116	126	121	116	126	1
241	122	116	126	122	117	126	0
242	123	117	127	122	117	126	1
243	123	117	128	123	118	127	0
244	124	118	128	123	118	127	1
245	125	118	129	124	119	128	1
246	125	119	129	124	119	128	1
247	126	119	130	125	120	129	1
248	126	120	130	125	120	129	1
249	127	120	131	126	121	130	1

Table F.1 (continued)

	RAPT Reading Composite			WRAT-4 Reading Composite			
		90% Confidence Interval			*90% Confidence Interval*		
Raw Score	DQ	Low	High	Standard Score	Low	High	**Discrepancy**
250	127	121	131	126	121	130	1
251	128	121	132	127	122	131	1
252	128	122	132	127	122	131	1
253	129	123	133	128	122	132	1
254	129	123	133	128	122	132	1
255	130	124	134	129	123	133	1
256	131	124	134	129	123	133	2
257	131	125	135	130	124	134	1
258	132	125	135	130	124	134	2
259	132	126	136	131	125	135	1
260	133	126	136	131	125	135	2
261	133	127	137	132	126	136	1
262	134	127	137	132	126	136	2
263	134	128	138	133	127	137	1
264	135	128	138	133	127	137	2
265	135	129	139	134	128	138	1
266	136	129	140	134	128	138	2
267	137	130	140	135	129	139	2
268	137	130	141	135	129	139	2
269	138	131	141	136	130	140	2
270	138	131	142	136	130	140	2
271	139	132	142	137	131	141	2
272	139	132	143	137	131	141	2
273	140	133	143	138	132	142	2
274	140	133	144	138	132	142	2
275	141	134	144	139	133	143	2
276	141	134	145	139	133	143	2
277	142	135	145	139	133	143	3
278	143	136	146	140	134	144	3
279	143	136	146	140	134	144	3
280	144	137	147	141	135	145	3
281	144	137	147	141	135	145	3
282	145	138	148	142	136	146	3
283	145	138	148	142	136	146	3
284	146	139	149	143	137	147	3
285	146	139	149	143	137	147	3
286	147	140	150	143	137	147	4
287	147	140	150	144	138	148	3
288	148	141	151	144	138	148	4
289	149	141	151	145	139	149	4
290	149	142	152	145	139	149	4

APPENDIX G:

Look Up Tables for WAIS-III Index Score

Equivalents of Sums of Scaled Scores

Calculated using RAPT methodology.

Table G. 1 Look-up Table FSIQ

Sum of Scaled Scores	FSIQ DQ	%ile Rank	90% Confidence Level	Sum of Scaled Scores	FSIQ DQ	%ile Rank	90% Confidence Level
11	39	0.0	37-44	111	101	52	97-104
12	40	0.0	37-45	112	101	53	98-105
13	40	0.0	38-45	113	102	55	98-105
14	41	0.0	38-46	114	102	57	99-106
15	41	0.0	39-46	115	103	58	99-107
16	42	0.0	40-47	116	104	60	100-107
17	43	0.0	40-48	117	104	61	101-108
18	43	0.0	41-48	118	105	63	101-109
19	44	0.0	41-49	119	106	64	102-109
20	44	0.0	42-49	120	106	66	102-110
21	45	0.0	43-50	121	107	67	103-110
22	46	0.0	43-51	122	107	69	104-111
23	46	0.0	44-51	123	108	70	104-112
24	47	0.0	44-52	124	109	72	105-112
25	48	0.0	45-52	125	109	73	105-113
26	48	0.0	46-53	126	110	74	106-113
27	49	0.0	46-54	127	110	76	107-114
28	49	0.0	47-54	128	111	77	107-115
29	50	0.0	47-55	129	112	78	108-115
30	51	0.0	48-55	130	112	79	108-116
31	51	0.1	49-56	131	113	81	109-116
32	52	0.1	49-57	132	114	82	110-117
33	52	0.1	50-57	133	114	83	110-118
34	53	0.1	50-58	134	115	84	111-118
35	54	0.1	51-58	135	115	85	111-119
36	54	0.1	52-59	136	116	86	112-119
37	55	0.1	52-60	137	117	87	113-120
38	56	0.2	53-60	138	117	88	113-121
39	56	0.2	53-61	139	118	88	114-121
40	57	0.2	54-61	140	119	89	114-122
41	57	0.2	55-62	141	119	90	115-122
42	58	0.3	55-63	142	120	91	116-123
43	59	0.3	56-63	143	120	91	116-124
44	59	0.3	57-64	144	121	92	117-124
45	60	0.4	57-64	145	122	93	117-125
46	60	0.4	58-65	146	122	93	118-125
47	61	0.5	58-66	147	123	94	119-126
48	62	0.5	59-66	148	123	94	119-127
49	62	0.6	60-67	149	124	95	120-127
50	63	0.7	60-68	150	125	95	120-128
51	64	0.8	61-68	151	125	95	121-128
52	64	0.9	61-69	152	126	96	122-129
53	65	1	62-69	153	127	96	122-130
54	65	1	63-70	154	127	96	123-130
55	66	1	63-71	155	128	97	123-131
56	67	1	64-71	156	128	97	124-131
57	67	1	64-72	157	129	97	125-132
58	68	2	65-72	158	130	98	125-133

Table G.1 (cont)

Sum of Scaled Scores	FSIQ DQ	%ile Rank	90% Confidence Level	Sum of Scaled Scores	FSIQ DQ	%ile Rank	90% Confidence Level
59	69	2	66-73	159	130	98	126-133
60	69	2	66-74	160	131	98	126-134
61	70	2	67-74	161	131	98	127-134
62	70	2	67-75	162	132	98	128-135
63	71	3	68-75	163	133	99	128-136
64	72	3	69-76	164	133	99	129-136
65	72	3	69-77	165	134	99	129-137
66	73	4	70-77	166	135	99	130-137
67	73	4	70-78	167	135	99	131-138
68	74	4	71-78	168	136	99.1	131-139
69	75	5	72-79	169	136	99.2	132-139
70	75	5	72-80	170	137	99.3	132-140
71	76	5	73-80	171	138	99.4	133-140
72	77	6	73-81	172	138	99.5	134-141
73	77	6	74-81	173	139	99.5	134-142
74	78	7	75-82	174	140	99.6	135-142
75	78	7	75-83	175	140	99.6	136-143
76	79	8	76-83	176	141	99.7	136-143
77	80	9	76-84	177	141	99.7	137-144
78	80	9	77-84	178	142	99.7	137-145
79	81	10	78-85	179	143	99.8	138-145
80	81	11	78-86	180	143	99.8	139-146
81	82	12	79-86	181	144	99.8	139-147
82	83	12	79-87	182	144	99.8	140-147
83	83	13	80-87	183	145	99.9	140-148
84	84	14	81-88	184	146	99.9	141-148
85	85	15	81-89	185	146	99.9	142-149
86	85	16	82-89	186	147	99.9	142-150
87	86	17	82-90	187	148	99.9	143-150
88	86	18	83-90	188	148	99.9	143-151
89	87	19	84-91	189	149	99.9	144-151
90	88	21	84-92	190	149	>99	145-152
91	88	22	85-92	191	150	>99	145-153
92	89	23	85-93	192	151	>99	146-153
93	90	24	86-93	193	151	>99	146-154
94	90	26	87-94	194	152	>99	147-154
95	91	27	87-95	195	152	>99	148-155
96	91	28	88-95	196	153	>99	148-156
97	92	30	88-96	197	154	>99	149-156
98	93	31	89-96	198	154	>99	149-157
99	93	33	90-97	199	155	>99	150-157
100	94	34	90-98	200	156	>99	151-158
101	94	36	91-98	201	156	>99	151-159
102	95	37	91-99	202	157	>99	152-159
103	96	39	92-99	203	157	>99	152-160
104	96	40	93-100	204	158	>99	153-160
105	97	42	93-101	205	159	>99	154-161
106	98	43	94-101	206	159	>99	154-162
107	98	45	95-102	207	160	>99	155-162
108	99	47	95-102	208	160	>99	155-163
109	99	48	96-103	209	161	>99	156-163
110	100	50	96-104				

Table G. 2 Look-up Table VIQ

Sum of Scaled Scores	VIQ DQ	%ile Rank	90% Confidence Level	Sum of Scaled Scores	VIQ DQ	%ile Rank	90% Confidence Level
6	45	0.0	42-50	61	101	53	97-105
7	46	0.0	43-51	62	102	55	98-106
8	47	0.0	44-52	63	103	58	99-107
9	48	0.0	45-53	64	104	61	100-108
10	49	0.0	46-54	65	105	63	101-109
11	50	0.0	47-55	66	106	66	102-110
12	51	0.1	48-56	67	107	68	103-111
13	52	0.1	49-57	68	108	71	104-112
14	53	0.1	50-58	69	109	73	105-113
15	54	0.1	51-59	70	110	75	106-114
16	55	0.1	52-60	71	111	77	107-115
17	56	0.2	53-61	72	112	79	108-116
18	57	0.2	54-62	73	113	81	109-117
19	58	0.3	55-63	74	114	83	110-118
20	59	0.3	56-64	75	115	85	111-119
21	60	0.4	57-65	76	116	86	112-120
22	61	0.5	58-66	77	117	88	113-121
23	62	0.6	59-67	78	118	89	114-122
24	63	0.7	60-68	79	119	90	115-123
25	64	0.8	61-69	80	120	91	116-124
26	65	1	62-70	81	122	92	117-125
27	66	1	63-71	82	123	93	118-126
28	67	1	64-72	83	124	94	119-127
29	68	2	65-73	84	125	95	120-128
30	69	2	66-74	85	126	96	121-129
31	70	2	67-75	86	127	96	122-130
32	71	3	68-76	87	128	97	123-131
33	72	3	69-77	88	129	97	124-132
34	73	4	70-78	89	130	98	125-133
35	74	4	71-79	90	131	98	126-134
36	75	5	72-80	91	132	98	127-135
37	76	6	73-81	92	133	99	128-136
38	77	7	74-82	93	134	99	129-137
39	78	8	75-83	94	135	99	130-138
40	80	9	76-84	95	136	99	131-139
41	81	10	77-85	96	137	99	132-140
42	82	11	78-86	97	138	99	133-141
43	83	12	79-87	98	139	99.5	134-142
44	84	14	80-88	99	140	99.6	135-143
45	85	15	81-89	100	141	99.7	136-144
46	86	17	82-90	101	142	99.7	137-145
47	87	19	83-91	102	143	99.8	138-146
48	88	21	84-92	103	144	99.8	139-147
49	89	23	85-93	104	145	99.9	140-148
50	90	25	86-94	105	146	99.9	141-149

Table G.2 (cont)

Sum of Scaled Scores	VIQ DQ	%ile Rank	90% Confidence Level	Sum of Scaled Scores	VIQ DQ	%ile Rank	90% Confidence Level
51	91	27	87-95	106	147	99.9	142-150
52	92	29	88-96	107	148	99.9	143-151
53	93	32	89-97	108	149	99.9	144-152
54	94	34	90-98	109	150	100.0	145-153
55	95	37	91-99	110	151	100.0	146-154
56	96	39	92-100	111	152	100.0	147-155
57	97	42	93-101	112	153	100.0	148-156
58	98	45	94-102	113	154	100.0	149-157
59	99	47	95-103	114	155	100.0	150-158
60	100	50	96-104				

Table G. 3 Look-up Table PIQ

Sum of Scaled Scores	PIQ DQ	%ile Rank	90% Confidence Level	Sum of Scaled Scores	PIQ DQ	%ile Rank	90% Confidence Level
5	40	0.0	38-50	53	104	60	98-110
6	42	0.0	39-51	54	105	64	99-111
7	43	0.0	40-52	55	107	67	100-112
8	44	0.0	42-53	56	108	70	102-113
9	46	0.0	43-55	57	109	73	103-115
10	47	0.0	44-56	58	111	76	104-116
11	48	0.0	45-57	59	112	79	105-117
12	50	0.0	47-58	60	113	81	107-118
13	51	0.1	48-60	61	115	83	108-120
14	52	0.1	49-61	62	116	86	109-121
15	54	0.1	50-62	63	117	88	110-122
16	55	0.1	52-63	64	119	89	112-123
17	56	0.2	53-65	65	120	91	113-125
18	58	0.2	54-66	66	121	92	114-126
19	59	0.3	55-67	67	123	93	115-127
20	60	0.4	57-68	68	124	94	117-128
21	61	0.5	58-70	69	125	95	118-130
22	63	0.7	59-71	70	127	96	119-131
23	64	0.8	60-72	71	128	97	120-132
24	65	1	62-73	72	129	97	122-133
25	67	1	63-75	73	131	98	123-135
26	68	2	64-76	74	132	98	124-136
27	69	2	65-77	75	133	98.8	125-137
28	71	3	67-78	76	135	98.9	127-138
29	72	3	68-80	77	136	99.2	128-140
30	73	4	69-81	78	137	99.3	129-141
31	75	5	70-82	79	139	99.5	130-142
32	76	6	72-83	80	140	99.6	132-143
33	77	7	73-85	81	141	99.7	133-145
34	79	8	74-86	82	142	99.8	134-146
35	80	9	75-87	83	144	99.8	135-147
36	81	11	77-88	84	145	99.9	137-148
37	83	12	78-90	85	146	99.9	138-150
38	84	14	79-91	86	148	99.9	139-151
39	85	17	80-92	87	149	99.9	140-152
40	87	19	82-93	88	150	100.0	142-153
41	88	21	83-95	89	152	100.0	143-155
42	89	24	84-96	90	153	100.0	144-156
43	91	27	85-97	91	154	100.0	145-157
44	92	30	87-98	92	156	100.0	147-158
45	93	33	88-100	93	157	100.0	148-160
46	95	36	89-101	94	158	100.0	149-161
47	96	40	90-102	95	160	100.0	150-162
48	97	43	92-103				
49	99	46	93-105				
50	100	50	94-106				
51	101	54	95-107				
52	103	57	97-108				

Table G. 4 Look-up Table VCI

Sum of Scaled Scores	VCI DQ	%ile Rank	90% Confidence Level	Sum of Scaled Scores	VCI DQ	%ile Rank	90% Confidence Level
3	51	0.0	48-57	43	124	94	118-128
4	52	0.1	49-59	44	126	96	120-129
5	54	0.1	51-61	45	127	97	122-131
6	56	0.2	53-63	46	129	97	123-133
7	58	0.3	55-64	47	131	98	125-135
8	60	0.4	57-66	48	133	99	127-136
9	62	0.5	58-68	49	135	99	129-138
10	63	0.7	60-70	50	137	99.3	130-140
11	65	1	62-71	51	138	99.5	132-142
12	67	1	64-73	52	140	99.6	134-143
13	69	2	65-75	53	142	99.7	136-145
14	71	3	67-77	54	144	99.8	137-147
15	73	3	69-78	55	146	99.9	139-149
16	74	4	71-80	56	148	99.9	141-151
17	76	6	72-82	57	149	100	143-152
18	78	7	74-84				
19	80	9	76-86				
20	82	11	78-87				
21	84	14	79-89				
22	85	16	81-91				
23	87	20	83-93				
24	89	23	85-94				
25	91	27	86-96				
26	93	31	88-98				
27	95	36	90-100				
28	96	40	92-101				
29	98	45	93-103				
30	100	50	95-105				
31	102	55	97-107				
32	104	60	99-108				
33	105	64	100-110				
34	107	69	102-112				
35	109	73	104-114				
36	111	77	106-115				
37	113	80	107-117				
38	115	84	109-119				
39	116	86	111-121				
40	118	89	113-122				
41	120	91	114-124				
42	122	93	116-126				

Table G. 5 Look-up Table POI

Sum of Scaled Scores	POI DQ	%ile Rank	90% Confidence Level
3	46	0.0	43-55
4	48	0.0	45-57
5	50	0.0	47-59
6	52	0.1	49-61
7	54	0.1	51-63
8	56	0.2	53-65
9	58	0.2	55-67
10	60	0.4	56-69
11	62	1	58-70
12	64	1	60-72
13	66	1	62-74
14	68	2	64-76
15	70	2	66-78
16	72	3	68-80
17	74	4	70-82
18	76	5	71-84
19	78	7	73-85
20	80	9	75-87
21	82	11	77-89
22	84	14	79-91
23	86	17	81-93
24	88	21	83-95
25	90	25	85-97
26	92	30	86-99
27	94	34	88-100
28	96	39	90-102
29	98	45	92-104
30	100	50	94-106
31	102	55	96-108
32	104	61	98-110
33	106	66	100-112
34	108	70	101-114
35	110	75	103-115
36	112	79	105-117
37	114	83	107-119
38	116	86	109-121
39	118	89	111-123
40	120	91	113-125
41	122	93	115-127
42	124	95	116-129
43	126	96	118-130
44	128	97	120-132
45	130	98	122-134
46	132	98	124-136
47	134	99	126-138
48	136	99	128-140
49	138	99	130-142
50	140	99.6	131-144
51	142	99.8	133-145
52	144	99.8	135-147
53	146	99.9	137-149
54	148	99.9	139-151
55	150	100.0	141-153
56	152	100.0	143-155
57	154	100.0	145-157

Table G. 6 Look-up Table WMI

Sum of Scaled Scores	WMI DQ	%ile Rank	90% Confidence Interval	Sum of Scaled Scores	WMI DQ	%ile Rank	90% Confidence Interval
3	46	0.0	44-56	48	136	99	128-140
4	48	0.0	45-57	49	138	99.4	129-142
5	50	0.0	47-59	50	140	99.6	131-143
6	52	0.1	49-61	51	142	99.7	133-145
7	54	0.1	51-63	52	144	99.8	135-147
8	56	0.2	53-65	53	146	99.9	137-149
9	58	0.3	55-67	54	148	99.9	139-151
10	60	0.4	57-69	55	150	>99	141-153
11	62	0.6	58-71	56	152	>99	143-155
12	64	1	60-72	57	154	>99	144-156
13	66	1	62-74				
14	68	2	64-76				
15	70	2	66-78				
16	72	3	68-80				
17	74	4	70-82				
18	76	6	72-84				
19	78	7	73-85				
20	80	9	75-87				
21	82	12	77-89				
22	84	14	79-91				
23	86	18	81-93				
24	88	21	83-95				
25	90	25	85-97				
26	92	30	87-99				
27	94	34	88-100				
28	96	40	90-102				
29	98	45	92-104				
30	100	50	94-106				
31	102	55	96-108				
32	104	60	98-110				
33	106	66	100-112				
34	108	70	101-113				
35	110	75	103-115				
36	112	79	105-117				
37	114	82	107-119				
38	116	86	109-121				
39	118	88	111-123				
40	120	91	113-125				
41	122	93	115-127				
42	124	94	116-128				
43	126	96	118-130				
44	128	97	120-132				
45	130	98	122-134				
46	132	98	124-136				
47	134	99	126-138				

Table G. 7 Look-up Table PSI

Sum of Scaled Scores	PSI DQ	%ile Rank	90% Confidence Interval
2	50	0.0	48-64
3	53	0.1	51-67
4	56	0.2	53-69
5	59	0.3	56-72
6	61	1	58-74
7	64	1	60-76
8	67	1	63-79
9	70	2	65-81
10	72	3	68-84
11	75	5	70-86
12	78	7	73-89
13	81	10	75-91
14	83	14	77-93
15	86	18	80-96
16	89	23	82-89
17	92	29	85-101
18	94	36	87-103
19	97	43	90-106
20	100	50	92-108
21	103	57	94-110
22	106	64	97-113
23	108	71	99-115
24	111	77	102-118
25	114	82	104-120
26	117	86	107-123
27	119	90	109-125
28	122	93	111-127
29	125	95	114-130
30	128	97	116-132
31	130	98	119-135
32	133	99	121-137
33	136	99.1	120-140
34	139	99.5	126-142
35	141	99.7	128-144
36	144	99.8	131-147
37	147	99.9	133-149
38	150	100.0	136-152

APPENDIX H:

Three Alternative Memory Composites using WMS-III subtest

Table H. 1 Scaled Scores, Means, Standard Deviations and Reliabilities for WMS-III subtests.

	Mean	SD	VAR	Reliability
LMI	10	3	9	.88
VPAI	10	3	9	.93
FI	10	3	9	.70
FPI	10	3	9	.83
VRI	10	3	9	.79
LMII	10	3	9	.79
VRII	10	3	9	.77

Table H. 2 Correlation Matrix for VCI and WMI subtests.

	LMI	VPAI	FI	FPI	VRI	LMII
LMI	1					
VPAI	.48	1				
FI	.14	.18	1			
FPI	0.40	.34	.30	1		
VRI	.39[a]	.39[a]		.35[a]	1	
LMII	.85				.42[c]	1
VRII	.33[b]				.62[d]	.38[c]

[a] Intercorrelations between VRI and LMI, VPI and FPI are the average of coefficients provided for three different age groups as follows:

Table H. 3 Intercorrelations between VRI and LMI, VPAI and FPI

	16-29	**30-64**	**65-89**	**Mean**
VRI with LMI	.39	.39	.38	.39
VRI with VPAI	.37	.44	.36	.39
VRI with FPI	.26	.35	.45	.35

[b] Intercorrelations between LMI and VRII are the average of coefficients provided for three different age groups as follows:

Table H. 4 Intercorrelations between LMI and VRI, VRII

	16-29	**30-64**	**65-89**	**Mean**
LMI with VRII	.38	.31	.31	.33

[c] Intercorrelations between LMII and VRI and VRII are the average of coefficients provided for three different age groups as follows:

Table H. 5 Intercorrelations between LMII and VRI, VRII

	16-29	**30-64**	**65-89**	**Mean**
LMII with VRI	.39	.40	.50	.42
LMII with VRI	.39	.37	.40	.38

[d] Intercorrelations between VRI and VRII are the average of coefficients provided for three different age groups as follows:

Table H. 6 Intercorrelations between VRI and VRII

	16-29	**30-64**	**65-89**	**Mean**
VRI with VRII	.59	.65	.62	.62

Figure H. 1 Model 1a Immediate Memory (LMI + VPAI + FP + FI)

Four Subtest Composite

NORMS:

	mean	sd	var	reliability	n
LMI	10	3	9	0.88	4
VPAI	10	3	9	0.93	
FI	10	3	9	0.7	
FPI	10	3	9	0.83	

INTERCORRELATIONS:

	VPAI	FI	FPI
LMI	0.48	0.14	0.4
VPAI		0.18	0.34
FI			0.3

COVARIANCE :

	VPAI	FI	FPI
LMI	8.64	2.52	7.2
VPAI		3.24	6.12
FI			5.4

COMPOSITE MEAN
40

COMPOSITE SD
8.313843876

SUM rxx	3.34
2 SUMrxy	3.68

COMPOSITE RELIABILITY
0.9140625

COMPOSITE SEp
6.083600321

COMPOSITE SEe
4.204076042

Table H. 7 Look-up Table Model 1a Immediate Memory (LMI + VPAI + FP + FI)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
4	35	0.00%	34	48	31	51
5	37	0.00%	35	49	32	52
6	39	0.00%	37	51	34	54
7	40	0.00%	39	52	36	56
8	42	0.01%	40	54	37	57
9	44	0.01%	42	56	39	59
10	46	0.02%	44	57	41	61
11	48	0.02%	45	59	42	62
12	49	0.04%	47	61	44	64
13	51	0.06%	49	62	45	65
14	53	0.09%	50	64	47	67
15	55	0.13%	52	66	49	69
16	57	0.19%	54	67	50	70
17	59	0.28%	55	69	52	72
18	60	0.41%	57	71	54	74
19	62	0.58%	58	72	55	75
20	64	0.81%	60	74	57	77
21	66	1.11%	62	76	59	79
22	68	1.52%	63	77	60	80
23	69	2.04%	65	79	62	82
24	71	2.71%	67	81	64	84
25	73	3.56%	68	82	65	85
26	75	4.61%	70	84	67	87
27	77	5.89%	72	85	69	89
28	78	7.45%	73	87	70	90
29	80	9.29%	75	89	72	92
30	82	11.45%	77	90	74	93
31	84	13.95%	78	92	75	95
32	86	16.80%	80	94	77	97
33	87	19.99%	82	95	78	98
34	89	23.52%	83	97	80	100
35	91	27.38%	85	99	82	102
36	93	31.52%	87	100	83	103
37	95	35.91%	88	102	85	105
38	96	40.49%	90	104	87	107
39	98	45.21%	91	105	88	108
40	100	50.00%	93	107	90	110
41	102	54.79%	95	109	92	112
42	104	59.51%	96	110	93	113
43	105	64.09%	98	112	95	115
44	107	68.48%	100	113	97	117
45	109	72.62%	101	115	98	118
46	111	76.48%	103	117	100	120
47	113	80.01%	105	118	102	122
48	114	83.20%	106	120	103	123
49	116	86.05%	108	122	105	125

Table H.7 (continued)

50	118	88.55%	110	123	107	126
51	120	90.71%	111	125	108	128
52	122	92.55%	113	127	110	130
53	123	94.11%	115	128	111	131
54	125	95.39%	116	130	113	133
55	127	96.44%	118	132	115	135
56	129	97.29%	119	133	116	136
57	131	97.96%	121	135	118	138
58	132	98.48%	123	137	120	140
59	134	98.89%	124	138	121	141
60	136	99.19%	126	140	123	143
61	138	99.42%	128	142	125	145
62	140	99.59%	129	143	126	146
63	141	99.72%	131	145	128	148
64	143	99.81%	133	146	130	150
65	145	99.87%	134	148	131	151
66	147	99.91%	136	150	133	153
67	149	99.94%	138	151	135	155
68	151	99.96%	139	153	136	156
69	152	99.98%	141	155	138	158
70	154	99.98%	143	156	139	159
71	156	99.99%	144	158	141	161
72	158	99.99%	146	160	143	163
73	160	100.00%	148	161	144	164
74	161	100.00%	149	163	146	166
75	163	100.00%	151	165	148	168
76	165	100.00%	152	166	149	169

Figure H. 2 Model 1a and 1b Auditory Memory (LMI + VPAI)

Two Subtest Composite

NORMS:

	mean	sd	var	reliability	n
LMI	10	3	9	0.88	2
VPAI	10	3	9	0.93	

INTERCORRELATIONS:

	VPAI
LMI	0.48

COMPOSITE MEAN

20

COMPOSITE SD

5.16139516

SUM rxx	1.81
2SUM rxy	0.96

COMPOSITE RELIABILITY

0.93581081

COMPOSITE SEe

3.67634477

COMPOSITE SEp

5.28753992

Table H. 8 Look-up Table Model 1a and 1b Auditory Memory (LMI + VPAI)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
2	48	0.02%	45	57	42	60
3	51	0.05%	48	60	45	62
4	54	0.10%	50	63	48	65
5	56	0.18%	53	65	50	68
6	59	0.33%	56	68	53	71
7	62	0.59%	59	71	56	73
8	65	1.00%	61	73	59	76
9	68	1.65%	64	76	61	79
10	71	2.63%	67	79	64	82
11	74	4.06%	69	82	67	84
12	77	6.06%	72	84	70	87
13	80	8.75%	75	87	72	90
14	83	12.25%	78	90	75	92
15	85	16.63%	80	92	78	95
16	88	21.92%	83	95	80	98
17	91	28.05%	86	98	83	101
18	94	34.92%	88	101	86	103
19	97	42.32%	91	103	89	106
20	100	50.00%	94	106	91	109
21	103	57.68%	97	109	94	111
22	106	65.08%	99	112	97	114
23	109	71.95%	102	114	99	117
24	112	78.08%	105	117	102	120
25	115	83.37%	108	120	105	122
26	117	87.75%	110	122	108	125
27	120	91.25%	113	125	110	128
28	123	93.94%	116	128	113	130
29	126	95.94%	118	131	116	133
30	129	97.37%	121	133	118	136
31	132	98.35%	124	136	121	139
32	135	99.00%	127	139	124	141
33	138	99.41%	129	141	127	144
34	141	99.67%	132	144	129	147
35	144	99.82%	135	147	132	150
36	146	99.90%	137	150	135	152
37	149	99.95%	140	152	138	155
38	152	99.98%	143	155	140	158

Figure H. 3 Model 1a Verbal Memory (FI +FPI)

Two Subtest Composite

NORMS:

	mean	sd	var	reliability	n
FI	10	3	9	0.7	2
FPI	10	3	9	0.83	

INTERCORRELATIONS:

	FPI
FI	0.3

COMPOSITE MEAN

20

COMPOSITE SD

4.83735465

SUM rxx	1.53
2SUM rxy	0.6

COMPOSITE RELIABILITY

0.81923077

COMPOSITE SEe

5.77240297

COMPOSITE SEp

8.60195984

Table H. 9 Look-up Table Model 1a Verbal Memory (FI +FPI)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
2	44	0.01%	45	64	40	68
3	47	0.02%	47	66	43	71
4	50	0.05%	50	69	45	74
5	53	0.10%	52	71	48	76
6	57	0.19%	55	74	50	79
7	60	0.36%	57	77	53	81
8	63	0.66%	60	79	55	84
9	66	1.15%	63	82	58	86
10	69	1.94%	65	84	60	89
11	72	3.14%	68	87	63	91
12	75	4.91%	70	89	65	94
13	78	7.39%	73	92	68	96
14	81	10.74%	75	94	71	99
15	84	15.07%	78	97	73	101
16	88	20.41%	80	99	76	104
17	91	26.76%	83	102	78	107
18	94	33.96%	85	104	81	109
19	97	41.81%	88	107	83	112
20	100	50.00%	90	110	86	114
21	103	58.19%	93	112	88	117
22	106	66.04%	96	115	91	119
23	109	73.24%	98	117	93	122
24	112	79.59%	101	120	96	124
25	116	84.93%	103	122	99	127
26	119	89.26%	106	125	101	129
27	122	92.61%	108	127	104	132
28	125	95.09%	111	130	106	135
29	128	96.86%	113	132	109	137
30	131	98.06%	116	135	111	140
31	134	98.85%	118	137	114	142
32	137	99.34%	121	140	116	145
33	140	99.64%	123	143	119	147
34	143	99.81%	126	145	121	150
35	147	99.90%	129	148	124	152
36	150	99.95%	131	150	126	155
37	153	99.98%	134	153	129	157
38	156	99.99%	136	155	132	160

Figure H. 4 Model 1b Immediate Memory (LMI + VPAI + FPI + VRI)

Four Subtest Composite

NORMS:

	mean	sd	var	reliability	n
LMI	10	3	9	0.88	4
VPAI	10	3	9	0.93	
FPI	10	3	9	0.83	
VRI	10	3	9	0.79	

INTERCORRELATIONS:

	VPAI	FPI	VRI
LMI	0.48	0.4	0.39
VPAI		0.34	0.39
FPI			0.35

COVARIANCE :

	VPAI	FPI	VRI
LMI	8.64	7.2	7.02
VPAI		6.12	7.02
FPI			6.3

COMPOSITE MEAN
40

COMPOSITE SD
8.848728722

SUM rxx	3.43
2 SUMrxy	4.7

COMPOSITE RELIABILITY
0.934482759

COMPOSITE SEp	
	5.340125852

COMPOSITE SEe
3.7115448

Table H. 10 Look-up Table Model 1b Immediate Memory (LMI + VPAI + FPI + VRI)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
4	39	0.00%	37	49	34	52
5	41	0.00%	38	51	36	53
6	42	0.01%	40	52	37	55
7	44	0.01%	42	54	39	56
8	46	0.01%	43	55	41	58
9	47	0.02%	45	57	42	60
10	49	0.03%	46	59	44	61
11	51	0.05%	48	60	45	63
12	53	0.08%	50	62	47	64
13	54	0.11%	51	63	48	66
14	56	0.17%	53	65	50	68
15	58	0.24%	54	66	52	69
16	59	0.33%	56	68	53	71
17	61	0.47%	57	70	55	72
18	63	0.65%	59	71	56	74
19	64	0.88%	61	73	58	75
20	66	1.19%	62	74	60	77
21	68	1.59%	64	76	61	79
22	69	2.10%	65	78	63	80
23	71	2.74%	67	79	64	82
24	73	3.53%	69	81	66	83
25	75	4.50%	70	82	67	85
26	76	5.68%	72	84	69	87
27	78	7.09%	73	85	71	88
28	80	8.75%	75	87	72	90
29	81	10.69%	76	89	74	91
30	83	12.92%	78	90	75	93
31	85	15.46%	80	92	77	95
32	86	18.30%	81	93	79	96
33	88	21.45%	83	95	80	98
34	90	24.89%	84	97	82	99
35	92	28.60%	86	98	83	101
36	93	32.56%	88	100	85	102
37	95	36.73%	89	101	86	104
38	97	41.06%	91	103	88	106
39	98	45.50%	92	105	90	107
40	100	50.00%	94	106	91	109
41	102	54.50%	95	108	93	110
42	103	58.94%	97	109	94	112
43	105	63.27%	99	111	96	114
44	107	67.44%	100	112	98	115
45	108	71.40%	102	114	99	117
46	110	75.11%	103	116	101	118
47	112	78.55%	105	117	102	120
48	114	81.70%	107	119	104	121
49	115	84.54%	108	120	105	123

Table H.10 (continued)

50	117	87.08%	110	122	107	125
51	119	89.31%	111	124	109	126
52	120	91.25%	113	125	110	128
53	122	92.91%	115	127	112	129
54	124	94.32%	116	128	113	131
55	125	95.50%	118	130	115	133
56	127	96.47%	119	131	117	134
57	129	97.26%	121	133	118	136
58	131	97.90%	122	135	120	137
59	132	98.41%	124	136	121	139
60	134	98.81%	126	138	123	140
61	136	99.12%	127	139	125	142
62	137	99.35%	129	141	126	144
63	139	99.53%	130	143	128	145
64	141	99.67%	132	144	129	147
65	142	99.76%	134	146	131	148
66	144	99.83%	135	147	132	150
67	146	99.89%	137	149	134	152
68	147	99.92%	138	150	136	153
69	149	99.95%	140	152	137	155
70	151	99.97%	141	154	139	156
71	153	99.98%	143	155	140	158
72	154	99.99%	145	157	142	159
73	156	99.99%	146	158	144	161
74	158	99.99%	148	160	145	163
75	159	100.00%	149	162	147	164
76	161	100.00%	151	163	148	166

Figure H. 5 Model 1b Visual Memory (FPI + VRI)

Two Subtest Composite

NORMS:

	mean	sd	var	reliability	n
VRI	10	3	9	0.79	2
FPI	10	3	9	0.83	

INTERCORRELATIONS:

	FPI
VRI	0.35

COMPOSITE MEAN

20

COMPOSITE SD

4.92950302

SUM rxx	1.62
2SUM rxy	0.7

COMPOSITE RELIABILITY

0.85925926

COMPOSITE SEe

5.21630871

COMPOSITE SEp

7.67310519

Table H. 11 Look-up Table Model 1b Visual Memory (FPI + VRI)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
2	45	0.01%	44	62	40	66
3	48	0.03%	47	64	43	68
4	51	0.06%	50	67	46	71
5	54	0.12%	52	69	48	73
6	57	0.23%	55	72	51	76
7	60	0.42%	57	75	53	79
8	63	0.75%	60	77	56	81
9	67	1.28%	63	80	59	84
10	70	2.12%	65	82	61	87
11	73	3.39%	68	85	64	89
12	76	5.23%	70	88	66	92
13	79	7.78%	73	90	69	94
14	82	11.18%	76	93	72	97
15	85	15.52%	78	96	74	100
16	88	20.86%	81	98	77	102
17	91	27.14%	84	101	79	105
18	94	34.25%	86	103	82	107
19	97	41.96%	89	106	85	110
20	100	50.00%	91	109	87	113
21	103	58.04%	94	111	90	115
22	106	65.75%	97	114	93	118
23	109	72.86%	99	116	95	121
24	112	79.14%	102	119	98	123
25	115	84.48%	104	122	100	126
26	118	88.82%	107	124	103	128
27	121	92.22%	110	127	106	131
28	124	94.77%	112	130	108	134
29	127	96.61%	115	132	111	136
30	130	97.88%	118	135	113	139
31	133	98.72%	120	137	116	141
32	137	99.25%	123	140	119	144
33	140	99.58%	125	143	121	147
34	143	99.77%	128	145	124	149
35	146	99.88%	131	148	127	152
36	149	99.94%	133	150	129	154
37	152	99.97%	136	153	132	157
38	155	99.99%	138	156	134	160

Figure H. 6 Model 2 Memory (LMI + LMII + VRI + VRII)

Four Subtest Composite

NORMS:

	mean	sd	var	reliability	n
LMI	10	3	9	0.88	4
LMII	10	3	9	0.79	
VRI	10	3	9	0.79	
VRII	10	3	9	0.77	

INTERCORRELATIONS:

	LMII	VRI	VRII
LMI	0.85	0.39	0.33
LMII		0.42	0.38
VRI			0.62

COVARIANCE :

	LMII	VRI	VRII
LMI	15.3	7.02	5.94
LMII		7.56	6.84
VRI			11.16

COMPOSITE MEAN
40

COMPOSITE SD
9.477341399

SUM rxx	3.23
2 SUMrxy	5.98

COMPOSITE RELIABILITY
0.922845691

COMPOSITE SEp
5.777548072

COMPOSITE SEe
4.002541978

Table H. 12 Look-up Table Model 2 Memory (LMI + LMII + VRI + VRII)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
4	43	0.01%	41	54	38	57
5	45	0.01%	42	55	39	58
6	46	0.02%	44	57	41	60
7	48	0.02%	45	58	42	61
8	49	0.04%	47	60	44	63
9	51	0.05%	48	61	45	64
10	53	0.08%	50	63	47	66
11	54	0.11%	51	64	48	67
12	56	0.16%	53	66	50	69
13	57	0.22%	54	67	51	70
14	59	0.30%	55	69	53	71
15	60	0.42%	57	70	54	73
16	62	0.57%	58	72	55	74
17	64	0.76%	60	73	57	76
18	65	1.01%	61	74	58	77
19	67	1.34%	63	76	60	79
20	68	1.74%	64	77	61	80
21	70	2.25%	66	79	63	82
22	72	2.88%	67	80	64	83
23	73	3.64%	69	82	66	85
24	75	4.57%	70	83	67	86
25	76	5.67%	72	85	69	88
26	78	6.98%	73	86	70	89
27	79	8.51%	74	88	72	90
28	81	10.27%	76	89	73	92
29	83	12.29%	77	90	74	93
30	84	14.57%	79	92	76	95
31	86	17.11%	80	93	77	96
32	87	19.93%	82	95	79	98
33	89	23.01%	83	96	80	99
34	91	26.33%	85	98	82	101
35	92	29.89%	86	99	83	102
36	94	33.65%	88	101	85	104
37	95	37.58%	89	102	86	105
38	97	41.64%	91	104	88	107
39	98	45.80%	92	105	89	108
40	100	50.00%	93	107	91	109
41	102	54.20%	95	108	92	111
42	103	58.36%	96	109	93	112
43	105	62.42%	98	111	95	114
44	106	66.35%	99	112	96	115
45	108	70.11%	101	114	98	117
46	109	73.67%	102	115	99	118
47	111	76.99%	104	117	101	120
48	113	80.07%	105	118	102	121
49	114	82.89%	107	120	104	123
50	116	85.43%	108	121	105	124

Table H.12 (continued)

51	117	87.71%	110	123	107	126
52	119	89.73%	111	124	108	127
53	121	91.49%	112	126	110	128
54	122	93.02%	114	127	111	130
55	124	94.33%	115	128	112	131
56	125	95.43%	117	130	114	133
57	127	96.36%	118	131	115	134
58	128	97.12%	120	133	117	136
59	130	97.75%	121	134	118	137
60	132	98.26%	123	136	120	139
61	133	98.66%	124	137	121	140
62	135	98.99%	126	139	123	142
63	136	99.24%	127	140	124	143
64	138	99.43%	128	142	126	145
65	140	99.58%	130	143	127	146
66	141	99.70%	131	145	129	147
67	143	99.78%	133	146	130	149
68	144	99.84%	134	147	131	150
69	146	99.89%	136	149	133	152
70	147	99.92%	137	150	134	153
71	149	99.95%	139	152	136	155
72	151	99.96%	140	153	137	156
73	152	99.98%	142	155	139	158
74	154	99.98%	143	156	140	159
75	155	99.99%	145	158	142	161
76	157	99.99%	146	159	143	162

Figure H. 7 Model 2 Auditory Memory (LMI + LMII)

Two Subtest Composite

NORMS:

	mean	sd	var	reliability	n
LMI	10	3	9	0.79	2
LMII	10	3	9	0.79	

INTERCORRELATIONS:

	LMII
LMI	0.85

COMPOSITE MEAN

20

COMPOSITE SD

5.77061522

SUM rxx	1.58
2SUM rxy	1.7

COMPOSITE RELIABILITY

0.88648649

COMPOSITE SEe

4.75829214

COMPOSITE SEp

6.94131721

Table H. 13 Look-up Table Model 2 Auditory (LMI + LMII)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
2	53	0.09%	51	66	47	70
3	56	0.16%	53	69	49	72
4	58	0.28%	55	71	52	75
5	61	0.47%	58	73	54	77
6	64	0.76%	60	76	56	79
7	66	1.21%	62	78	59	81
8	69	1.88%	64	80	61	84
9	71	2.83%	67	83	63	86
10	74	4.16%	69	85	66	88
11	77	5.94%	71	87	68	91
12	79	8.28%	74	89	70	93
13	82	11.26%	76	92	72	95
14	84	14.92%	78	94	75	98
15	87	19.31%	81	96	77	100
16	90	24.41%	83	99	79	102
17	92	30.16%	85	101	82	105
18	95	36.45%	88	103	84	107
19	97	43.12%	90	106	86	109
20	100	50.00%	92	108	89	111
21	103	56.88%	94	110	91	114
22	105	63.55%	97	112	93	116
23	108	69.84%	99	115	95	118
24	110	75.59%	101	117	98	121
25	113	80.69%	104	119	100	123
26	116	85.08%	106	122	102	125
27	118	88.74%	108	124	105	128
28	121	91.72%	111	126	107	130
29	123	94.06%	113	129	109	132
30	126	95.84%	115	131	112	134
31	129	97.17%	117	133	114	137
32	131	98.12%	120	136	116	139
33	134	98.79%	122	138	119	141
34	136	99.24%	124	140	121	144
35	139	99.53%	127	142	123	146
36	142	99.72%	129	145	125	148
37	144	99.84%	131	147	128	151
38	147	99.91%	134	149	130	153

Figure H. 8 Model 2 Verbal Memory (VRI + VRII)

Two Subtest Composite

NORMS:

	mean	sd	var	reliability	n
VRI	10	3	9	0.79	2
VRII	10	3	9	0.77	

INTERCORRELATIONS:

	VRII
VRI	0.62

COMPOSITE MEAN

20

COMPOSITE SD

5.4

SUM rxx	1.56
2SUM rxy	1.24

COMPOSITE RELIABILITY

0.86419753

COMPOSITE SEe

5.13868034

COMPOSITE SEp

7.54729032

Table H. 14 Look-up Table Model 2 Verbal Memory (VRI + VRII)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
2	50	0.04%	48	65	44	69
3	53	0.08%	51	68	47	72
4	56	0.15%	53	70	49	74
5	58	0.27%	56	72	52	76
6	61	0.48%	58	75	54	79
7	64	0.80%	60	77	56	81
8	67	1.31%	63	80	59	84
9	69	2.08%	65	82	61	86
10	72	3.20%	68	84	64	88
11	75	4.78%	70	87	66	91
12	78	6.92%	72	89	68	93
13	81	9.74%	75	92	71	96
14	83	13.33%	77	94	73	98
15	86	17.72%	80	96	76	100
16	89	22.94%	82	99	78	103
17	92	28.93%	84	101	80	105
18	94	35.56%	87	104	83	108
19	97	42.65%	89	106	85	110
20	100	50.00%	92	108	88	112
21	103	57.35%	94	111	90	115
22	106	64.44%	96	113	92	117
23	108	71.07%	99	116	95	120
24	111	77.06%	101	118	97	122
25	114	82.28%	104	120	100	124
26	117	86.67%	106	123	102	127
27	119	90.26%	108	125	104	129
28	122	93.08%	111	128	107	132
29	125	95.22%	113	130	109	134
30	128	96.80%	116	132	112	136
31	131	97.92%	118	135	114	139
32	133	98.69%	120	137	116	141
33	136	99.20%	123	140	119	144
34	139	99.52%	125	142	121	146
35	142	99.73%	128	144	124	148
36	144	99.85%	130	147	126	151
37	147	99.92%	132	149	128	153
38	150	99.96%	135	152	131	156

Appendix I:

Three Alternative Cognitive Composites using WAIS-III subtest

Table I. 1 Scaled Scores, Means, Standard Deviations and Reliabilities for WAIS-III subtests.

	Mean	SD	VAR	Reliability
Information (INF)	10	3	9	.91
Vocabulary (VOC)	10	3	9	.93
Communication (COM)	10	3	9	.84
Similarities (SIM)	10	3	9	.86
Picture Arrangement (PA)	10	3	9	.74
Matrix Reasoning (MR)	10	3	9	.90
Block Design (BD)	10	3	9	.86
Object Assembly (OA)	10	3	9	.70
Arithmetic (AR)	10	3	9	.88
Picture Completion (PC)	10	3	9	.83
Digit Span (DSp)	10	3	9	.90
Letter Number Sequencing (LNS)	10	3	9	.82
Symbol Search (SS)	10	3	9	.77
Digit Symbol (DSy)	10	3	9	.84

* Reliability coefficients based on the average coefficient calculated with Fisher's z transformation (Wechsler, 1997, p.50).

Table I. 2 Correlation Matrix for WAIS-III Subtests

	INF	VOC	COM	SIM	PA	MR	BD	OA	AR	PC	DSp	LNS	SS	DSy
INF	1													
VOC	.77	1												
COM	.70	.75	1											
SIM	.70	.76	.70	1										
PA	.54	.53	.50	.52	1									
MR	.53	.54	.52	.54	.50	1								
BD	.48	.50	.49	.52	.49	.60	1							
OA	.40	.44	.45	.47	.46	.49	.61	1						
AR	.63	.60	.57	.57	.44	.58	.54	.39	1					
PC	.46	.47	.46	.48	.49	.48	.52	.52	.40	1				
DSp	.40	.45	.39	.40	.33	.42	.36	.26	.52	.30	1			
LNS	.47	.50	.44	.46	.39	.47	.43	.29	.55	.41	.57	1		
SS	.46	.48	.44	.48	.45	.48	.53	.47	.52	.49	.41	.49	1	
DSy	.38	.44	.37	.40	.37	.40	.41	.33	.43	.39	.36	.44	.65	1

Figure I. 1 Crystallized Intelligence (INF + VOC + COM + SIM+PA)

Five Subtest Composite

NORMS:

	mean	sd	var	reliability	n
INF	10	3	9	0.91	5
VOC	10	3	9	0.93	
COM	10	3	9	0.84	
SIM	10	3	9	0.86	
PA	10	3	9	0.74	

INTERCORRELATIONS:

	VOC	COM	SIM	PA
INF	0.77	0.7	0.7	0.54
VOC		0.75	0.76	0.53
COM			0.7	0.5
SIM				0.52

COVARIANCE :

	VOC	COM	SIM	PA
INF	13.86	12.6	12.6	9.72
VOC		13.5	13.68	9.54
COM			12.6	9
SIM				9.36

COMPOSITE MEAN	
50	
COMPOSITE SD	
12.70669115	
SUM rxx	4.28
2*SUM rxy	12.94
COMPOSITE RELIABILITY	
0.959866221	
COMPOSITE SEe	
2.944093796	
COMPOSITE SEp	
4.206873972	

Table I. 3 Look-up Table Crystallized Intelligence (INF + VOC + COM + SIM+PA)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
5	47	0.0%	44	54	42	56
6	48	0.0%	45	55	43	57
7	49	0.0%	46	56	44	58
8	50	0.0%	48	57	46	59
9	52	0.1%	49	58	47	60
10	53	0.1%	50	60	48	62
11	54	0.1%	51	61	49	63
12	55	0.1%	52	62	50	64
13	56	0.2%	53	63	51	65
14	58	0.2%	54	64	52	66
15	59	0.3%	56	65	53	67
16	60	0.4%	57	66	55	68
17	61	0.5%	58	67	56	70
18	62	0.6%	59	69	57	71
19	63	0.7%	60	70	58	72
20	65	0.9%	61	71	59	73
21	66	1.1%	62	72	60	74
22	67	1.4%	63	73	61	75
23	68	1.7%	65	74	63	76
24	69	2%	66	75	64	77
25	70	2%	67	77	65	79
26	72	3%	68	78	66	80
27	73	4%	69	79	67	81
28	74	4%	70	80	68	82
29	75	5%	71	81	69	83
30	76	6%	73	82	70	84
31	78	7%	74	83	72	85
32	79	8%	75	84	73	87
33	80	9%	76	86	74	88
34	81	10%	77	87	75	89
35	82	12%	78	88	76	90
36	83	14%	79	89	77	91
37	85	15%	80	90	78	92
38	86	17%	82	91	80	93
39	87	19%	83	92	81	94
40	88	22%	84	93	82	96
41	89	24%	85	95	83	97
42	91	26%	86	96	84	98
43	92	29%	87	97	85	99
44	93	32%	88	98	86	100
45	94	35%	90	99	87	101
46	95	38%	91	100	89	102
47	96	41%	92	101	90	103
48	98	44%	93	103	91	105
49	99	47%	94	104	92	106
50	100	50%	95	105	93	107
51	101	53%	96	106	94	108
52	102	56%	97	107	95	109
53	104	59%	99	108	97	110
54	105	62%	100	109	98	111
55	106	65%	101	110	99	113
56	107	68%	102	112	100	114
57	108	71%	103	113	101	115
58	109	74%	104	114	102	116
59	111	76%	105	115	103	117

Table I.3 (continued)

60	112	78%	107	116	104	118
61	113	81%	108	117	106	119
62	114	83%	109	118	107	120
63	115	85%	110	120	108	122
64	117	86%	111	121	109	123
65	118	88%	112	122	110	124
66	119	90%	113	123	111	125
67	120	91%	114	124	112	126
68	121	92%	116	125	113	127
69	122	93%	117	126	115	128
70	124	94%	118	127	116	130
71	125	95%	119	129	117	131
72	126	96%	120	130	118	132
73	127	96%	121	131	119	133
74	128	97%	122	132	120	134
75	130	98%	123	133	121	135
76	131	98.0%	125	134	123	136
77	132	98.3%	126	135	124	137
78	133	98.6%	127	137	125	139
79	134	98.9%	128	138	126	140
80	135	99.1%	129	139	127	141
81	137	99.3%	130	140	128	142
82	138	99.4%	131	141	129	143
83	139	99.5%	133	142	130	144
84	140	99.6%	134	143	132	145
85	141	99.7%	135	144	133	147
86	142	99.8%	136	146	134	148
87	144	99.8%	137	147	135	149
88	145	99.9%	138	148	136	150
89	146	99.9%	139	149	137	151
90	147	99.9%	140	150	138	152
91	148	99.9%	142	151	140	153
92	150	100.0%	143	152	141	154
93	151	100.0%	144	154	142	156
94	152	100.0%	145	155	143	157
95	153	100.0%	146	156	144	158

Figure I. 2 Fluid Intelligence (MR + BD + OA + SIM+PA+AR)

Six Subtest Composite

NORMS:

	mean	sd	var	reliability	n
MR	10	3	9	0.9	6
BD	10	3	9	0.86	
OA	10	3	9	0.7	
SIM	10	3	9	0.86	
PA	10	3	9	0.74	
AR	10	3	9	0.88	

INTERCORRELATIONS:

	BD	OA	SIM	PA	AR
MR	0.76	0.6	0.45	0.77	0.75
BD		0.57	0.4	0.7	0.7
OA			0.52	0.63	0.57
SIM				0.4	0.39
PA					0.7

COVARIANCE :

	BD	OA	SIM	PA	AR
MR	13.68	10.8	8.1	13.86	13.5
BD		10.26	7.2	12.6	12.6
OA			9.36	11.34	10.26
SIM				7.2	7.02
PA					12.6

COMPOSITE MEAN	
60	
COMPOSITE SD	
14.641721	
SUM rxx	4.94
SUM rxy	8.91
2 SUMrxy	17.82
COMPOSITE RELIABILITY	
0.9554996	
COMPOSITE SEe	
3.0930616	
COMPOSITE SEp	
4.4248869	

Table I. 4 Look-up Table Fluid Intelligence (MR + BD + OA + SIM+PA+AR)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
6	45	0.0%	42	52	40	54
7	46	0.0%	43	53	41	55
8	47	0.0%	44	54	42	56
9	48	0.0%	45	55	43	57
10	49	0.0%	46	56	44	58
11	50	0.0%	47	57	45	59
12	51	0.1%	48	58	46	60
13	52	0.1%	49	59	47	61
14	53	0.1%	50	60	48	62
15	54	0.1%	51	61	49	63
16	55	0.1%	52	62	50	64
17	56	0.2%	53	63	51	65
18	57	0.2%	54	64	52	66
19	58	0.3%	55	65	53	67
20	59	0.3%	56	66	54	68
21	60	0.4%	57	67	55	69
22	61	0.5%	58	68	56	70
23	62	0.6%	59	69	57	71
24	63	0.7%	60	70	58	72
25	64	0.8%	61	71	58	73
26	65	1%	62	72	59	74
27	66	1%	63	73	60	75
28	67	1%	64	74	61	76
29	68	2%	65	75	62	77
30	69	2%	66	76	63	78
31	70	2%	67	77	64	79
32	71	3%	68	78	65	80
33	72	3%	68	79	66	81
34	73	4%	69	80	67	82
35	74	4%	70	81	68	83
36	75	5%	71	82	69	84
37	76	6%	72	83	70	85
38	77	7%	73	84	71	86
39	78	8%	74	85	72	87
40	80	9%	75	85	73	88
41	81	10%	76	86	74	89
42	82	11%	77	87	75	90
43	83	12%	78	88	76	91
44	84	14%	79	89	77	92
45	85	15%	80	90	78	93
46	86	17%	81	91	79	94
47	87	19%	82	92	80	95
48	88	21%	83	93	81	96
49	89	23%	84	94	82	96
50	90	25%	85	95	83	97
51	91	27%	86	96	84	98
52	92	29%	87	97	85	99
53	93	32%	88	98	86	100
54	94	34%	89	99	87	101
55	95	37%	90	100	88	102
56	96	39%	91	101	89	103
57	97	42%	92	102	90	104
58	98	45%	93	103	91	105
59	99	47%	94	104	92	106
60	100	50%	95	105	93	107

Table I.4 (continued)

61	101	53%	96	106	94	108
62	102	55%	97	107	95	109
63	103	58%	98	108	96	110
64	104	61%	99	109	97	111
65	105	63%	100	110	98	112
66	106	66%	101	111	99	113
67	107	68%	102	112	100	114
68	108	71%	103	113	101	115
69	109	73%	104	114	102	116
70	110	75%	105	115	103	117
71	111	77%	106	116	104	118
72	112	79%	107	117	104	119
73	113	81%	108	118	105	120
74	114	83%	109	119	106	121
75	115	85%	110	120	107	122
76	116	86%	111	121	108	123
77	117	88%	112	122	109	124
78	118	89%	113	123	110	125
79	119	90%	114	124	111	126
80	120	91%	115	125	112	127
81	122	92%	115	126	113	128
82	123	93%	116	127	114	129
83	124	94%	117	128	115	130
84	125	95%	118	129	116	131
85	126	96%	119	130	117	132
86	127	96%	120	131	118	133
87	128	97%	121	132	119	134
88	129	97%	122	132	120	135
89	130	98%	123	133	121	136
90	131	98%	124	134	122	137
91	132	98%	125	135	123	138
92	133	99%	126	136	124	139
93	134	99%	127	137	125	140
94	135	99%	128	138	126	141
95	136	99%	129	139	127	142
96	137	99%	130	140	128	142
97	138	99%	131	141	129	143
98	139	99.5%	132	142	130	144
99	140	99.6%	133	143	131	145
100	141	99.7%	134	144	132	146
101	142	99.7%	135	145	133	147
102	143	99.8%	136	146	134	148
103	144	99.8%	137	147	135	149
104	145	99.9%	138	148	136	150
105	146	99.9%	139	149	137	151
106	147	99.9%	140	150	138	152
107	148	99.9%	141	151	139	153
108	149	99.9%	142	152	140	154
109	150	100.0%	143	153	141	155
110	151	100.0%	144	154	142	156
111	152	100.0%	145	155	143	157
112	153	100.0%	146	156	144	158
113	154	100.0%	147	157	145	159
114	155	100.0%	148	158	146	160

Figure I. 3 Broad Visualization (PC + BD + OA + MR)

Four Subtest Composite

NORMS:

	mean	sd	var	reliability	n
PC	10	3	9	0.83	4
BD	10	3	9	0.86	
OA	10	3	9	0.7	
MR	10	3	9	0.9	

INTERCORRELATIONS:

	BD	OA	MR
PC	0.52	0.52	0.48
BD		0.61	0.6
OA			0.49

COVARIANCE :

	BD	OA	MR
PC	9.36	9.36	8.64
BD		10.98	10.8
OA			8.82

COMPOSITE MEAN	
40	
COMPOSITE SD	
9.693296653	
SUM rxx	3.29
2 SUMrxy	6.44
COMPOSITE RELIABILITY	
0.931992337	
COMPOSITE SEe	
3.776385791	
COMPOSITE SEp	
5.437169648	

Table I. 5 Look- up Table Broad Visualization (PC+BD+OA+MR)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
4	44	0.01%	42	54	39	57
5	46	0.02%	43	56	41	58
6	47	0.02%	45	57	42	60
7	49	0.03%	46	59	43	61
8	50	0.05%	48	60	45	63
9	52	0.07%	49	61	46	64
10	54	0.10%	51	63	48	66
11	55	0.14%	52	64	49	67
12	57	0.19%	53	66	51	69
13	58	0.27%	55	67	52	70
14	60	0.37%	56	69	54	71
15	61	0.50%	58	70	55	73
16	63	0.66%	59	72	56	74
17	64	0.88%	61	73	58	76
18	66	1.16%	62	74	59	77
19	68	1.51%	64	76	61	79
20	69	1.95%	65	77	62	80
21	71	2.50%	66	79	64	82
22	72	3.17%	68	80	65	83
23	74	3.97%	69	82	67	84
24	75	4.94%	71	83	68	86
25	77	6.09%	72	85	69	87
26	78	7.43%	74	86	71	89
27	80	8.99%	75	87	72	90
28	81	10.79%	77	89	74	92
29	83	12.82%	78	90	75	93
30	85	15.11%	79	92	77	94
31	86	17.66%	81	93	78	96
32	88	20.46%	82	95	80	97
33	89	23.51%	84	96	81	99
34	91	26.80%	85	98	82	100
35	92	30.30%	87	99	84	102
36	94	33.99%	88	100	85	103
37	95	37.85%	89	102	87	105
38	97	41.83%	91	103	88	106
39	98	45.89%	92	105	90	107
40	100	50.00%	94	106	91	109
41	102	54.11%	95	108	93	110
42	103	58.17%	97	109	94	112
43	105	62.15%	98	111	95	113
44	106	66.01%	100	112	97	115
45	108	69.70%	101	113	98	116
46	109	73.20%	102	115	100	118
47	111	76.49%	104	116	101	119
48	112	79.54%	105	118	103	120
49	114	82.34%	107	119	104	122
50	115	84.89%	108	121	106	123
51	117	87.18%	110	122	107	125
52	119	89.21%	111	123	108	126
53	120	91.01%	113	125	110	128
54	122	92.57%	114	126	111	129
55	123	93.91%	115	128	113	131
56	125	95.06%	117	129	114	132
57	126	96.03%	118	131	116	133
58	128	96.83%	120	132	117	135

Table I.5 (continued)

59	129	97.50%	121	134	118	136
60	131	98.05%	123	135	120	138
61	132	98.49%	124	136	121	139
62	134	98.84%	126	138	123	141
63	136	99.12%	127	139	124	142
64	137	99.34%	128	141	126	144
65	139	99.50%	130	142	127	145
66	140	99.63%	131	144	129	146
67	142	99.73%	133	145	130	148
68	143	99.81%	134	147	131	149
69	145	99.86%	136	148	133	151
70	146	99.90%	137	149	134	152
71	148	99.93%	139	151	136	154
72	150	99.95%	140	152	137	155
73	151	99.97%	141	154	139	157
74	153	99.98%	143	155	140	158
75	154	99.98%	144	157	142	159
76	156	99.99%	146	158	143	161

Figure I. 4 Short-Term Memory (AR+DSp+LNS)

Three Subtest Composite

NORMS:

	mean	sd	var	reliability	n
AR	10	3	9	0.88	3
DSp	10	3	9	0.9	
LNS	10	3	9	0.82	

INTERCORRELATIONS:

	DSp	LNS
AR	0.52	0.55
DSp		0.57

COVARIANCE:

	DSp	LNS
AR	9.36	9.9
DSp		10.26

COMPOSITE MEAN	**30**
COMPOSITE SD	**7.517978452**
SUM rxx	2.6
2*SUM rxy	3.28
COMPOSITE RELIABILITY	**0.936305732**
COMPOSITE SEe	**3.663112646**
COMPOSITE SEp	**5.267789332**

Table I. 6 Look-up Table Short-Term Memory (AR+DSp+LNS)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
3	46	0.0%	44	56	41	58
4	48	0.0%	45	57	43	60
5	50	0.0%	47	59	45	62
6	52	0.1%	49	61	47	64
7	54	0.1%	51	63	48	66
8	56	0.2%	53	65	50	68
9	58	0.3%	55	67	52	69
10	60	0.4%	57	69	54	71
11	62	1%	58	71	56	73
12	64	1%	60	72	58	75
13	66	1%	62	74	60	77
14	68	2%	64	76	61	79
15	70	2%	66	78	63	81
16	72	3%	68	80	65	82
17	74	4%	70	82	67	84
18	76	6%	72	84	69	86
19	78	7%	73	85	71	88
20	80	9%	75	87	73	90
21	82	12%	77	89	75	92
22	84	14%	79	91	76	94
23	86	18%	81	93	78	96
24	88	21%	83	95	80	97
25	90	25%	85	97	82	99
26	92	30%	87	99	84	101
27	94	34%	88	100	86	103
28	96	40%	90	102	88	105
29	98	45%	92	104	89	107
30	100	50%	94	106	91	109
31	102	55%	96	108	93	111
32	104	60%	98	110	95	112
33	106	66%	100	112	97	114
34	108	70%	101	113	99	116
35	110	75%	103	115	101	118
36	112	79%	105	117	103	120
37	114	82%	107	119	104	122
38	116	86%	109	121	106	124
39	118	88%	111	123	108	125
40	120	91%	113	125	110	127
41	122	93%	115	127	112	129
42	124	94%	116	128	114	131
43	126	96%	118	130	116	133
44	128	97%	120	132	118	135
45	130	98%	122	134	119	137
46	132	98%	124	136	121	139
47	134	99%	126	138	123	140
48	136	99%	128	140	125	142
49	138	99%	129	142	127	144
50	140	99.6%	131	143	129	146
51	142	99.7%	133	145	131	148
52	144	99.8%	135	147	132	150
53	146	99.9%	137	149	134	152
54	148	99.9%	139	151	136	153
55	150	100.0%	141	153	138	155
56	152	100.0%	143	155	140	157
57	154	100.0%	144	156	142	159

Figure I. 5 Broad Speediness (DSy+SS+OA)

Three Subtest Composite

NORMS:

	mean	sd	var	reliability	n
Dsy	10	3	9	0.84	3
SS	10	3	9	0.77	
OA	10	3	9	0.7	

INTERCORRELATIONS:

	SS	OA
Dsy	0.65	0.33
SS		0.47

COVARIANCE :

	SS	OA
Dsy	11.7	5.94
SS		8.46

COMPOSITE MEAN

30

COMPOSITE SD

7.28697468

SUM rxx	2.31
2*SUM rxy	2.9

COMPOSITE RELIABILITY

0.88305085

COMPOSITE SEe

4.8203953

COMPOSITE SEp

7.03915976

Table I. 7Look-up Table Broad Speediness (DSy+SS+OA)

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
3	44	0.0%	43	59	39	62
4	46	0.0%	45	61	41	64
5	49	0.0%	47	62	43	66
6	51	0.0%	48	64	45	68
7	53	0.1%	50	66	47	70
8	55	0.1%	52	68	48	72
9	57	0.2%	54	70	50	73
10	59	0.3%	56	72	52	75
11	61	0%	58	73	54	77
12	63	1%	59	75	56	79
13	65	1%	61	77	58	81
14	67	1%	63	79	59	82
15	69	2%	65	81	61	84
16	71	3%	67	82	63	86
17	73	4%	68	84	65	88
18	75	5%	70	86	67	90
19	77	7%	72	88	68	92
20	79	8%	74	90	70	93
21	81	11%	76	92	72	95
22	84	14%	78	93	74	97
23	86	17%	79	95	76	99
24	88	21%	81	97	78	101
25	90	25%	83	99	79	102
26	92	29%	85	101	81	104
27	94	34%	87	102	83	106
28	96	39%	88	104	85	108
29	98	45%	90	106	87	110
30	100	50%	92	108	88	112
31	102	55%	94	110	90	113
32	104	61%	96	112	92	115
33	106	66%	98	113	94	117
34	108	71%	99	115	96	119
35	110	75%	101	117	98	121
36	112	79%	103	119	99	122
37	114	83%	105	121	101	124
38	116	86%	107	122	103	126
39	119	89%	108	124	105	128
40	121	92%	110	126	107	130
41	123	93%	112	128	108	132
42	125	95%	114	130	110	133
43	127	96%	116	132	112	135
44	129	97%	118	133	114	137
45	131	98%	119	135	116	139
46	133	99%	121	137	118	141
47	135	99%	123	139	119	142
48	137	99%	125	141	121	144
49	139	100%	127	142	123	146
50	141	99.7%	128	144	125	148
51	143	99.8%	130	146	127	150
52	145	99.9%	132	148	128	152
53	147	99.9%	134	150	130	153
54	149	100.0%	136	152	132	155
55	151	100.0%	138	153	134	157
56	154	100.0%	139	155	136	159
57	156	100.0%	141	157	138	161

Appendix J:

Alternative Battery Structure Composite Calculations

Figure J. 1 Work Knowledge (WAIS3-VO + WAIS3-SI + STW + WRAT3-Reading)

Four Subtest Composite

NORMS:

	mean	sd	var	reliability	n
WAIS3-VO	10	3	9	0.93	4
WAIS3-SI	10	3	9	0.86	
STW	10	3	9	0.83	
WRAT3-Reading	10	3	9	0.95	

INTERCORRELATIONS:

	WAIS3-SI	STW	WRAT3-Reading
WAIS3-VO	0.591	0.534	0.666
WAIS3-SI		0.381	0.423
STW			0.611

COVARIANCE :

	WAIS3-SI	STW	WRAT3-Reading
WAIS3-VO	10.638	9.612	11.988
WAIS3-SI		6.858	7.614
STW			10.998

COMPOSITE MEAN	
40	
COMPOSITE SD	
9.680289252	
SUM rxx	3.57
2 SUMrxy	6.412
COMPOSITE RELIABILITY	
0.958701498	
COMPOSITE SEe	
2.984696045	
COMPOSITE SEp	
4.266213001	

Table J. 1 Look-Up Table Work Knowledge

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
4	44	0.01%	42	51	40	54
5	46	0.01%	43	53	41	55
6	47	0.02%	45	54	42	56
7	49	0.03%	46	56	44	58
8	50	0.05%	48	57	45	59
9	52	0.07%	49	59	47	61
10	54	0.10%	51	60	48	62
11	55	0.14%	52	62	50	64
12	57	0.19%	54	63	51	65
13	58	0.26%	55	65	53	67
14	60	0.36%	56	66	54	68
15	61	0.49%	58	68	56	70
16	63	0.66%	59	69	57	71
17	64	0.88%	61	71	59	73
18	66	1.15%	62	72	60	74
19	67	1.50%	64	74	62	76
20	69	1.94%	65	75	63	77
21	71	2.48%	67	77	65	79
22	72	3.15%	68	78	66	80
23	74	3.95%	70	80	68	82
24	75	4.92%	71	81	69	83
25	77	6.06%	73	83	71	85
26	78	7.41%	74	84	72	86
27	80	8.96%	76	86	74	88
28	81	10.76%	77	87	75	89
29	83	12.79%	79	89	77	91
30	85	15.08%	80	90	78	92
31	86	17.63%	82	92	80	94
32	88	20.43%	83	93	81	95
33	89	23.48%	85	94	83	97
34	91	26.77%	86	96	84	98
35	92	30.27%	88	97	86	100
36	94	33.97%	89	99	87	101
37	95	37.83%	91	100	89	103
38	97	41.82%	92	102	90	104
39	98	45.89%	94	103	92	106
40	100	50.00%	95	105	93	107
41	102	54.11%	97	106	94	108
42	103	58.18%	98	108	96	110
43	105	62.17%	100	109	97	111
44	106	66.03%	101	111	99	113
45	108	69.73%	103	112	100	114
46	109	73.23%	104	114	102	116
47	111	76.52%	106	115	103	117
48	112	79.57%	107	117	105	119
49	114	82.37%	108	118	106	120
50	115	84.92%	110	120	108	122
51	117	87.21%	111	121	109	123
52	119	89.24%	113	123	111	125
53	120	91.04%	114	124	112	126
54	122	92.59%	116	126	114	128
55	123	93.94%	117	127	115	129
56	125	95.08%	119	129	117	131
57	126	96.05%	120	130	118	132
58	128	96.85%	122	132	120	134

Table J.1 (continued)

59	129	97.52%	123	133	121	135
60	131	98.06%	125	135	123	137
61	133	98.50%	126	136	124	138
62	134	98.85%	128	138	126	140
63	136	99.12%	129	139	127	141
64	137	99.34%	131	141	129	143
65	139	99.51%	132	142	130	144
66	140	99.64%	134	144	132	146
67	142	99.74%	135	145	133	147
68	143	99.81%	137	146	135	149
69	145	99.86%	138	148	136	150
70	146	99.90%	140	149	138	152
71	148	99.93%	141	151	139	153
72	150	99.95%	143	152	141	155
73	151	99.97%	144	154	142	156
74	153	99.98%	146	155	144	158
75	154	99.99%	147	157	145	159
76	156	99.99%	149	158	146	160

Figure J. 2 Processing Speed (SDMT-W + SDMT-O + TMT-A)

Three Subtest Composite

NORMS:

	mean	sd	var	reliability	n
SDMT-W	10	3	9	0.8	3
SDMT-O	10	3	9	0.76	
TMT-A	10	3	9	0.89	

INTERCORRELATIONS:

	SDMT-O	TMT-A
SDMT-W	0.853	0.518
SDMT-O		0.449

COVARIANCE :

	SDMT-O	TMT-A
SDMT-W	15.354	9.324
SDMT-O		8.082

COMPOSITE MEAN
30

COMPOSITE SD
7.730459236

SUM rxx	2.45
2*SUM rxy	3.64

COMPOSITE RELIABILITY
0.917168675

COMPOSITE See
4.134406462

COMPOSITE SEp
5.977488184

Table J. 2 Look-Up Table Processing Speed

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
3	48	0.0%	45	59	42	62
4	50	0.0%	47	61	44	64
5	51	0.1%	49	62	46	65
6	53	0.1%	51	64	47	67
7	55	0.1%	52	66	49	69
8	57	0.2%	54	68	51	71
9	59	0.3%	56	69	53	72
10	61	0.5%	58	71	55	74
11	63	1%	59	73	56	76
12	65	1%	61	75	58	78
13	67	1%	63	77	60	80
14	69	2%	65	78	62	81
15	71	3%	67	80	64	83
16	73	4%	68	82	65	85
17	75	5%	70	84	67	87
18	77	6%	72	85	69	88
19	79	8%	74	87	71	90
20	81	10%	75	89	72	92
21	83	12%	77	91	74	94
22	84	15%	79	93	76	96
23	86	18%	81	94	78	97
24	88	22%	83	96	80	99
25	90	26%	84	98	81	101
26	92	30%	86	100	83	103
27	94	35%	88	101	85	104
28	96	40%	90	103	87	106
29	98	45%	91	105	88	108
30	100	50%	93	107	90	110
31	102	55%	95	109	92	112
32	104	60%	97	110	94	113
33	106	65%	99	112	96	115
34	108	70%	100	114	97	117
35	110	74%	102	116	99	119
36	112	78%	104	117	101	120
37	114	82%	106	119	103	122
38	116	85%	107	121	104	124
39	117	88%	109	123	106	126
40	119	90%	111	125	108	128
41	121	92%	113	126	110	129
42	123	94%	115	128	112	131
43	125	95%	116	130	113	133
44	127	96%	118	132	115	135
45	129	97%	120	133	117	136
46	131	98%	122	135	119	138
47	133	99%	123	137	120	140
48	135	99%	125	139	122	142
49	137	99%	127	141	124	144
50	139	99.5%	129	142	126	145
51	141	99.7%	131	144	128	147
52	143	99.8%	132	146	129	149
53	145	99.9%	134	148	131	151
54	147	99.9%	136	149	133	153
55	149	99.9%	138	151	135	154
56	150	100.0%	139	153	136	156
57	152	100.0%	141	155	138	158

Figure J. 3 Verbal Fluency (COWAT + ANIMALS)

Two Subtest Composite

NORMS:

	mean	sd	var	reliability	n
COWAT	10	3	9	0.82	2
ANIMALS	10	3	9	0.55	

INTERCORRELATIONS:

	ANIMALS
COWAT	0.402

COMPOSITE MEAN

20

COMPOSITE SD

5.023544565

SUM rxx	1.37
2SUM rxy	0.804

COMPOSITE RELIABILITY

0.77532097

COMPOSITE SEe

6.260561618

COMPOSITE SEp

9.47351115

Table J. 3 Look-up Table Verbal Fluency

			90% Test Confidence		90% Retest Confidence	
SumSS	**DQ**	**%ile**	**Low**	**High**	**Low**	**High**
2	46	0.02%	48	69	43	74
3	49	0.04%	50	71	45	76
4	52	0.07%	53	73	47	79
5	55	0.14%	55	76	50	81
6	58	0.27%	57	78	52	83
7	61	0.48%	60	80	54	86
8	64	0.85%	62	83	57	88
9	67	1.43%	64	85	59	90
10	70	2.33%	67	87	61	92
11	73	3.66%	69	89	64	95
12	76	5.56%	71	92	66	97
13	79	8.17%	73	94	68	99
14	82	11.62%	76	96	70	102
15	85	15.98%	78	99	73	104
16	88	21.29%	80	101	75	106
17	91	27.52%	83	103	77	109
18	94	34.53%	85	106	80	111
19	97	42.11%	87	108	82	113
20	100	50.00%	90	110	84	116
21	103	57.89%	92	113	87	118
22	106	65.47%	94	115	89	120
23	109	72.48%	97	117	91	123
24	112	78.71%	99	120	94	125
25	115	84.02%	101	122	96	127
26	118	88.38%	104	124	98	130
27	121	91.83%	106	127	101	132
28	124	94.44%	108	129	103	134
29	127	96.34%	111	131	105	136
30	130	97.67%	113	133	108	139
31	133	98.57%	115	136	110	141
32	136	99.15%	117	138	112	143
33	139	99.52%	120	140	114	146
34	142	99.73%	122	143	117	148
35	145	99.86%	124	145	119	150
36	148	99.93%	127	147	121	153
37	151	99.96%	129	150	124	155
38	154	99.98%	131	152	126	157

Figure J. 4 Intercorrelations between WK and PS

Composite 1 with Composite 2

Intercorrelations

			Composite j				
			Subtest 1	Subtest 2	Subtest 3	Subtest 4	Subtest 5
			SDMT-W	SDMT-O	TMT-A		
Composite i		SD	3	3	3		
Subtest 1	WRAT3-Reading	3	0.2	0.157	0.061		
Subtest 2	WAIS3-VO	3	0.109	0.025	0.011		
Subtest 3	WAIS3-SI	3	0.101	0.054	0.073		
Subtest 4	STW	3	0.066	0.066	0.012		
Subtest 5							
Subtest 6							

Covariance: SDi*SDj*rij

			Composite j				
			Subtest 1	Subtest 2	Subtest 3	Subtest 4	Subtest 5
			SDMT-W	SDMT-O	TMT-A	0	0
Composite i		SD	3	3	3	0	0
Subtest 1	WRAT3-Reading	3	1.8	1.413	0.549	0	0
Subtest 2	WAIS3-VO	3	0.981	0.225	0.099	0	0
Subtest 3	WAIS3-SI	3	0.909	0.486	0.657	0	0
Subtest 4	STW	3	0.594	0.594	0.108	0	0
Subtest 5	0	0	0	0	0	0	0
Subtest 6	0	0	0	0	0	0	0

	r	SD		
Composite i:	0.959	9.68	Σ(SDi*SDj*rij)	8.415
Composite j:	0.917	7.731	SDi*SDj	74.83608

Composite Intercorrelation: 0.11

Significance	
p<.05	p<.01
10.35	13.63

Difference	Abnormality		
15	10%	5%	1%
22.6%	26	33	46

Figure J. 5 Intercorrelations between WK and VF

Composite 1 with Composite 2

Intercorrelations

			Composite j				
			Subtest 1	Subtest 2	Subtest 3	Subtest 4	Subtest 5
			COWAT	ANIMALS			
Composite i		SD	3	3			
Subtest 1	WRAT3-Reading	3	0.445	0.253			
Subtest 2	WAIS3-VO	3	0.335	0.232			
Subtest 3	WAIS3-SI	3	0.271	0.277			
Subtest 4	STW	3	0.357	0.259			
Subtest 5							
Subtest 6							

Covariance: SDi*SDj*rij

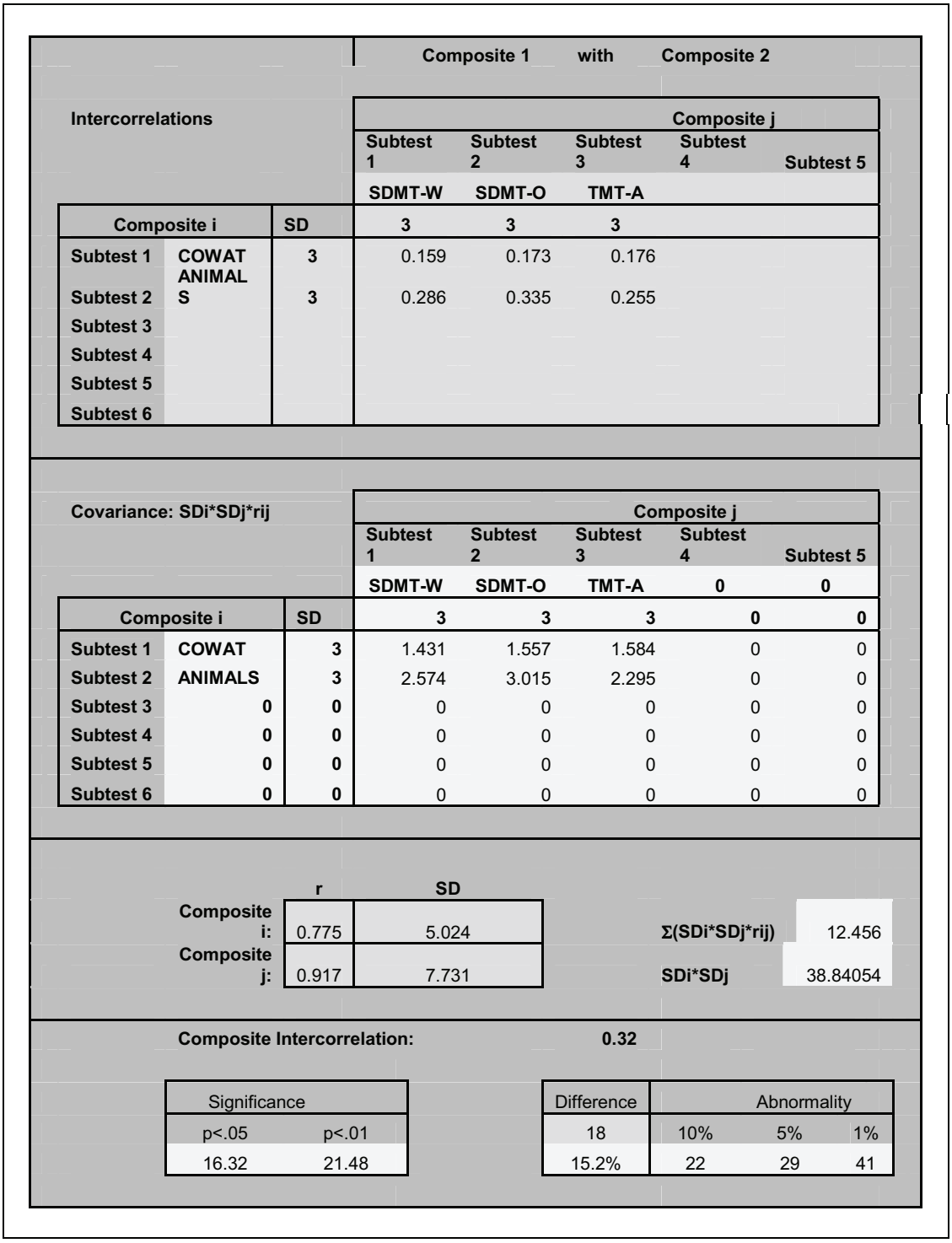
			Composite j				
			Subtest 1	Subtest 2	Subtest 3	Subtest 4	Subtest 5
			COWAT	ANIMALS	0	0	0
Composite i		SD	3	3	0	0	0
Subtest 1	WRAT3-Reading	3	4.005	2.277	0	0	0
Subtest 2	WAIS3-VO	3	3.015	2.088	0	0	0
Subtest 3	WAIS3-SI	3	2.439	2.493	0	0	0
Subtest 4	STW	3	3.213	2.331	0	0	0
Subtest 5	0	0	0	0	0	0	0
Subtest 6	0	0	0	0	0	0	0

	r	SD		
Composite i:	0.959	9.68	Σ(SDi*SDj*rij)	21.861
Composite j:	0.775	5.024	SDi*SDj	48.63232

Composite Intercorrelation: 0.45

Significance		Difference	Abnormality		
p<.05	p<.01	33	10%	5%	1%
15.16	19.96	1.8%	20	26	37

Figure J. 6 Intercorrelations between VF and PS



Composite 1 with Composite 2

Intercorrelations

			Composite j				
			Subtest 1	Subtest 2	Subtest 3	Subtest 4	Subtest 5
			SDMT-W	SDMT-O	TMT-A		
Composite i		SD	3	3	3		
Subtest 1	COWAT	3	0.159	0.173	0.176		
Subtest 2	ANIMALS	3	0.286	0.335	0.255		
Subtest 3							
Subtest 4							
Subtest 5							
Subtest 6							

Covariance: SDi*SDj*rij

			Composite j				
			Subtest 1	Subtest 2	Subtest 3	Subtest 4	Subtest 5
			SDMT-W	SDMT-O	TMT-A	0	0
Composite i		SD	3	3	3	0	0
Subtest 1	COWAT	3	1.431	1.557	1.584	0	0
Subtest 2	ANIMALS	3	2.574	3.015	2.295	0	0
Subtest 3	0	0	0	0	0	0	0
Subtest 4	0	0	0	0	0	0	0
Subtest 5	0	0	0	0	0	0	0
Subtest 6	0	0	0	0	0	0	0

	r	SD		
Composite i:	0.775	5.024	Σ(SDi*SDj*rij)	12.456
Composite j:	0.917	7.731	SDi*SDj	38.84054

Composite Intercorrelation: **0.32**

Significance	
p<.05	p<.01
16.32	21.48

Difference	Abnormality		
18	10%	5%	1%
15.2%	22	29	41