Application of Item Response Theory to Measures of Verbal Learning

Indrani K. Thiruselvam
Marquette University

APPLICATION OF ITEM RESPONSE THEORY TO MEASURES OF VERBAL LEARNING

by

Indrani K Thiruselvam, M.A.

A Dissertation submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy

Milwaukee, Wisconsin

August 2015

ABSTRACT
APPLICATION OF ITEM RESPONSE THEORY TO MEASURES OF VERBAL LEARNING

Indrani K Thiruselvam

Marquette University, 2015

This study utilized item response theory (IRT) methods to investigate if item parameters of select trials in the California Verbal Learning Test-Second Edition (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000) and the Logical Memory subtests of the Wechsler Memory Scales – Fourth Edition (WMS-IV; Wechsler, 2009) vary as a function of the serial position effect. In addition, this study compared the effectiveness of CVLT-II and LM in quantifying verbal memory functioning, and determined if a weighted scoring approach improves the quantification of verbal memory. Archival data from 755 individuals (516 college students, 239 patients at a neuropsychology clinic) were utilized in this study. The serial position effect was only evident in Trials 1 and 5 of the CVLT-II. CVLT-II trials were more effective than LM trials in quantifying verbal memory, although LM trials had, on average, higher difficulty levels. The weighted scoring approach utilized in this study did not lead to improvements in the quantification of verbal memory. Nevertheless, findings indicate that some items or trials perform better than others in discriminating between examinees with low levels of memory ability, and that it is important to more closely evaluate item properties of tests used in clinical decision-making.

ACKNOWLEDGEMENTS

Indrani K Thiruselvam

I would like to express my deepest gratitude to my mentor and dissertation committee chair, Dr. James Hoelzle, for his guidance throughout this dissertation and my graduate career. His insights, patience, and humor were instrumental in getting me to the finish line.

I would also like to thank my dissertation committee members, Drs. Kristy Nielson, Anthony Porcelli, and Jennifer Geiger for their time and effort in providing invaluable comments and asking stimulating questions that guided the formation, refinement, and completion of this study. Drs. Jennifer Geiger and Sara Swanson were very helpful in coordinating efforts to obtain archival data from Froedtert Hospital/Medical College of Wisconsin. I also want to acknowledge Applied Psychological Measurement Inc for the grant that contributed to the completion of this project and Wen Zheng for her invaluable insights on statistical analyses.

This study would not have been completed without the assistance of members of the Hoelzle lab. Graduate students Elisabeth Vogt and Erin Quasney provided encouragement and helped coordinate data collection throughout this study. Undergraduate students Jennifer Albrecht, Lauren Benjamin, Sean Brown, Megan Drury, Lauren Gallagher, Abigail Helbling, Megan Herodes, Katheryn Meek, Elle Merrill, Bethany Quering, Rachel Rechner, Kathryn Ritchie, Ashley Schuett, Kara Signorelli, Kristin Skells, Madeline Stollfus, Janel Wasisco, and Caleb Zutavern provided immense

help with data collection and data entry. They could not have done a more vibrant job in bringing this project to fruition.

My beloved Marquette University Department of Psychology cohort was my source of support throughout this dissertation, providing listening ears, helping hands, and generous hearts. They made graduate school so fun! I also want to thank Hugo Pereira and my wonderful network of friends for their unwavering encouragement, care, and patience.

Finally, this dissertation would not have been possible without the tremendous support from family. I am so appreciative of the empowerment and blessings that my parents offered in my pursuit of cultural and educational exchange half way across the world from Malaysia. I am humbled by the selflessness, love, and kindness from Carol and Jon Reese, Nellie Peterman, Pam and Dave Uhrig, and Cindy and Russ Hum, my amazing families away from home.

TABLE OF CONTENTS

LIST OF TABLES

RESULTS

DISCUSSION

LIST OF FIGURES

INTRODUCTION

RESULTS

DISCUSSION

APPENDIX A

INTRODUCTION

Psychological Assessments

Formal psychological assessment is a unique aspect of psychological practice that is not provided by other health care providers (Meyer et al., 2001). Assessment involves gathering and integrating related data to make an evaluation. It typically involves the use of different data-collection methods, such as tests, interviews, behavioral observations, case studies, and special measurement procedures (Cohen & Swerdlik, 2010). Tests are a large component of assessment, and can be defined as systematic procedures for observing and describing behavior in numerical or categorical systems (Cronbach, 1984). A closely related concept is measurement, which involves assigning numbers to objects so that attributes are accurately represented as numerical properties (Krantz, Luce, Suppers, & Tversky, 1971).

One of the key assumptions in assessment is that psychological traits and states can be measured by considering the types of item content (and behaviors) that would be indicative of the targeted construct (Cohen & Swerdlik, 2010). For example, the construct of memory can be measured by performance on a test that requires an individual to encode and retain material. However, measurement of psychological constructs is not precise and there is acknowledgement that error is part of the assessment process. Despite the existence of error variance, tests and assessments are beneficial in that they aid many forms of decision-making. In the present day, assessments are used to describe current functioning, such as severity of a disturbance, or capacity for independent living, and to confirm, refute, or modify a clinician’s impression of clients (Meyer et al., 2001). The assessment process can help identify therapeutic needs and offer guidance on possible interventions and likely outcomes; assessment tools can provide insight on issues likely to surface throughout the treatment process, and factors that may aid or hinder progress in treatment, such as psychological mindedness, family support, and personality characteristics.

Psychological assessments focus on various domains of functioning. Personality and intellectual assessments have always been major components of clinical psychology (Garfield, 1974; Wade & Baker, 1977) and have been prominent components of training in clinical psychology (Piotrowski & Zalewski, 1993). Neuropsychological assessment, which evaluates a broader range of cognitive abilities than intellectual assessment, is a more specific domain in psychology. This research will investigate the psychometric properties and utility of commonly-used neuropsychological measures, specifically the California Verbal Learning Test-Second Edition (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000) and the Logical Memory subtests of the Wechsler Memory Scales – Fourth Edition (WMS-IV; Wechsler, 2009).

Neuropsychological assessments: An overview. Broadly, neuropsychology is the study of brain-behavior relationships, whereas clinical neuropsychology is the "study of brain behavior relationships, and the clinical application of that knowledge to human problems" (American Psychological Association Division 40, 2012). The primary purpose of a neuropsychological assessment is "to draw inferences about the structural and functional characteristics of a person's brain by evaluating an individual's behavior in defined stimulus-response situations" (Benton, 1994, p. 1). Clinical neuropsychologists utilize an assortment of tests to detect cerebral dysfunction in situations in which there is no clear anatomic evidence of brain abnormalities, or to document the implication of an identifiable lesion (Lezak, Howieson, & Loring, 2004). Various abilities are assessed during a neuropsychological evaluation, including attention, verbal memory, executive functions, visuospatial skills, nonverbal memory, and intelligence, among others (Rabin, Barr, & Burton, 2005). Neuropsychological tests have been shown to have a moderate level of ecological validity in predicting everyday cognitive functioning, particularly when it involves predicting functioning in specific domains to which the neuropsychological tests correspond (Chaytor & Schmitter-Edgecombe, 2003; Makatura et al., 1999). Given that memory is one of the most frequently assessed domains of neuropsychological functioning (Rabin, Barr, & Burton, 2005; Lees-Haley, Smith, Willias, & Dunn, 1996), the focus of this paper will be on the domain of memory and how it is quantified.

Overview of Memory Function and Assessment. Various neuropsychological measures exists for quantifying and assessing memory, given that there are many aspects to memory. Memory is described as a complex function by which an organism registers, stores, retains, and retrieves previous exposure to an experience or event (Emilien, Durlach, Antoniadis, van der Linden, & Maloteaux, 2004). Memory is the outcome of learning, or the process of acquiring new information, and develops when something is learned, either through single or repeated exposures to information (Gazzaniga, Ivry, & Mangun, 2008). It is arguably among the most important brain faculty, given that very few cognitive processes, such as recognition, planning, language, decision-making, and creativity can function effectively without contributions from memory (Tranel & Damasio, 2002).

Different forms of memory exist. At the broadest level, memory can be divided into sensory, short-term, and long term memory (Baddeley, 2002). Sensory memory has a lifetime spanning milliseconds to seconds. Short-term memory spans seconds to minutes, whereas long-term memory spans days to years. Evidence for these separate memory systems was obtained through contrasts of patients with different neurological issues. For example, those with classic amnestic syndrome, with damage to the temporal lobes and hippocampi, typically demonstrate difficulties in learning and remembering new material, although their sensory and short term memory is relatively intact (Milner, 1966). On the other hand, those with damage to the left perisylvian cortex demonstrate limited short-term memory but seemingly normal long-term memory (Shallice & Warrington, 1970).

Several models have been developed to explain the interaction between these memory systems. Some researchers argue that short-term memory is necessary in the formation of long-term memory. Atkinson and Shiffrin (1968) proposed that information from the environment flows through sensory memory (part of the perceptual system) into a limited capacity short-term memory store. The likelihood that information transfers from short-term to long-term memory is dependent on factors such as duration during which information resides in short-term memory (Atkinson & Shiffrin, 1968), and depth to which an item is processed (Craik & Lockhart, 1972). These variables (capacity of short-term memory, duration in short-term memory, and depth to which information is processed) can explain various memory phenomena, including the serial position effect, which is the observation that people better recall items located at the beginning and end of the list, compared to items in the middle of the list (Deese & Kaufman, 1957).

Other researchers argue that although there is an interaction between short- and long-term memory, the former is not necessary for long-term learning to occur. Long-term memory can be divided into declarative or explicit memory, and nondeclarative or implicit memory (Gazzaniga, Ivry, & Mangun, 2008). The focus of this paper will be on declarative memory, as this component is most frequently evaluated in neuropsychological assessments. Declarative memory involves recollection of events (episodic memory) and facts (semantic memory). Nondeclarative memory involves procedural memory (motor and cognitive skills), the perceptual representation system (priming), classical conditioning, and nonassociative learning (habituation, sensitization). Evidence for this distinction between declarative and nondeclarative long-term memory is found in observations that densely amnesic patients are able to learn motor skills (Corkin, 1968), and action-consequence associations (Claparede, 1911). These studies demonstrate that learning does not require retrieval of the original learning episode but can be based on implicit memory accessed indirectly through performance.

A number of brain regions have been identified to underlie these different aspects of memory. The integrity of these structures are inferred from performances on various memory tests. The medial temporal lobe is primarily associated with declarative memory and consists of the hippocampal region and the adjacent entorhinal, perirhinal, and parahippocampal cortices (Squire, Stark, & Clark, 2004). This system accesses and influences signals from the entire brain; it creates records of interactions with the external environment, as well as internal thought processes (Tranel & Damasio, 2002). A lesion to the medial temporal lobe results in profound forgetfulness (Levy et al., 2003; Milner, 1971a). The associated impairment may be multimodal because the medial temporal lobe is one of the cortical processing association areas that receive input from all sensory modalities (Lavenex & Amaral, 2000). In addition, the two components of declarative memory, semantic and episodic memory can be distinguished by their sensitivity to brain damage (Vargha-Khadem et al., 1997). Specifically, episodic memory depends primarily on the hippocampus, whereas semantic memory depends primarily on the entorhinal, perirhinal, and parahippocampal cortices.

Neuropsychological measures have been developed with the specific aim of measuring these different aspects of memory and are used to infer the functioning of brain areas primarily associated with these subdomains. Assessment of memory functioning usually begins with an evaluation of a patient's orientation for time, place, and person (Wilson, 2002). Performance on forward digit span and recency effects on list-learning tests are indicative of verbal short-term memory, whereas performance on the token test, spatial span, and symbol span are considered to be indicative of visual short-term memory. Long-term verbal episodic memory is assessed using story recall or list-learning tests, whereas, remote episodic memory is typically assessed informally through semi-structured interviews involving autobiographical information. Long-term visual episodic memory is assessed with visual reproduction tests. Semantic memory can be assessed through category fluency tests, naming tests, and general knowledge questions. Nondeclarative memory can be assessed through visual motor tracking, mirror tracing, and stem completion tasks. Given that many different measures are available to test similar constructs of memory, it is important to evaluate the utility and psychometric properties of these measures to determine which is best suited for use in a given context. It is essential to utilize measures that have demonstrated the highest reliability and validity to increase the precision of inferences made about the integrity of brain structures and functioning within a given domain.

Psychometrics in Clinical Assessment Tools: How Clinicians Evaluate the Instruments Used

Key psychometric concepts. Psychological tests attempt to measure one or several hypothetical constructs that are typically unobservable, known as latent trait (Baker & Kim, 2004). Examples of latent traits include intelligence, arithmetic ability, depression, or memory. In measuring latent traits, it is essential that tests demonstrate consistency in measurement, or reliability. Conceptually, reliability pertains to the proportion of the total variance in an obtained score that is due to true variance (variance in the measured construct), as opposed to error variance (Cohen & Swerdlik, 2010; Schmidt & Embretson, 2013). Practically, reliability pertains to test consistency and exists in various forms. Test-retest consistency is the degree to which pairs of scores from the same people on two different administrations of the same test are correlated. Parallel- and alternate-form reliability estimates involve comparing two parallel (means and variances of observed test scores are equal) or alternate test forms to determine the degree to which they correlate with each other. Internal consistency pertains to the degree to which items within a test are correlated with each other, either by comparing two halves of a single test (split-half reliability) or by comparing each item with every other item on a test or scale (inter-item correlation, item total correlation, Cronbach's alpha).

Evaluating the reliability of an instrument is essential in determining whether a measure is psychometrically sound. Once it is established that measurement is reliable, it is necessary to evaluate the validity of the instrument. Validity pertains to the meaningfulness of a test score and the manner in which a score is interpreted; it is an estimate of the degree to which a test measures what it purports to measure (Cohen & Swerdlik, 2010). A valid test is necessarily reliable; conversely, a reliable test is not necessarily valid. For example, a digit forward test can produce consistent scores over time, but it may not be a valid measure of math ability.

The Standards for Educational and Psychological Testing (AERA, 1999) identify a number of sources of validity evidence. Evidence based on test content is obtained from the relationship between test content (themes, format of items, questions on test, wording, administration and scoring procedures) and the intended construct measured. For example, how relevantly and adequately do test items measure the gamut of behaviors intended to be sampled? Evidence based on response processes involves analyzing test-takers' responses to determine the fit between the construct and nature of performance. An example could include documenting aspects of performances relevant to a given construct, such as order of words recalled and intrusive errors in a verbal list learning task, or eye movement, in a measure of attention. Other examples include taking into account the degree to which observers or judges appropriately record and evaluate data according to the intended test interpretation. Evidence based on internal structure involves the degree to which relationships among test items and components conform to the construct or theory on which the test is based. In other words, how appropriate are inferences drawn about a test-taker's standings on a construct, based on test scores? Also, do scores on tests vary as theoretically predicted? Evidence based on relations to other variables pertains to the degree to which relationships with external variables are consistent with the construct underlying proposed test interpretations. For example, how strongly does the test correlate with other measures of related constructs, and how weakly does the test correlate with other tests that are not theoretically related to the construct measured? Finally, evidence based on the consequences of testing pertains to the degree to which some benefit will be realized from the intended use of scores.

The reliability of a test impacts the standard error of measurement (SEM), which is “an estimate of the variability expected for observed scores when the true score is held constant” (Dudek, 1979, p. 335). This variability in observed scores occurs because few tests are perfectly reliable. The larger the SEM, the lower the reliability. Therefore, much effort is taken to minimize error because it results in a large SEM, which leads to lower test precision and questionable test validity. For example, the Logical Memory (LM) Immediate Recall subtest has a lower internal consistency reliability (average r = .82) than the LM Delayed Recall subtest (average r = .85). Consequently, the LM I subtest has a larger SEM (1.28) than the LM II subtest (SEM = 1.17). The reliability value used to derive the SEM can be selected depending on the inferences made; if inferences about individual scores are to be made concerning what the score might be if the examinee was retested at a later date, it is logical to use test-retest reliability, whereas if inferences about individual scores are to be made about scores on two different forms of test, then it would be logical to use an alternate form reliability estimate (Harvill, 1991). Nevertheless, SEM is typically estimated based on internal consistency reliability (Slick, 2004).

The SEM is used to generate confidence intervals, which is the range of scores that is likely to contain the true score (Cohen & Swerdlik, 2010). For example, if a test has a SEM of 2 and a patient obtained an observed standard score of 90 on a memory test, we can be 68% certain that her score lies between 88 and 92 (90 ± 1 SEM), or we can be 96% confident that her true score lies between 86 and 94 (90 ± 2 SEM).

Apart from the SEM, which is tied to the inherent imperfections of a given test, there are other sources of error in obtained test scores. These include examinee factors (e.g., fatigue, lack of motivation, reactivity to the testing situation, and guessing), examiner factors (e.g., test administration and rapport-building skills, scoring and interpretation mistakes), and environmental factors (e.g., room noise level, lighting, and temperature). A memory test administered to an unmotivated examinee in a noisy room is unlikely to provide an accurate reflection of the examinee's memory functioning. It is assumed that the examinee is putting forth sufficient effort in an environment where extraneous variables (i.e., those not pertaining to memory) are minimized. Such errors could decrease the reliability of a test and result in test bias. Flaugher (1978) identified biases that could also arise from different interpretations of tests, over-interpretations of the meaning of test scores, sexism, test content, selection models, wrong criterion, and testing atmosphere.

Differential item functioning (DIF) is another form of bias in test scores, whereby the probability of endorsing an item is higher for one group than the other, across various trait levels (Swaminathan & Rogers, 1990). In other words, despite two people from different groups having the same latent trait level, they have a different probability of obtaining a correct score on a given item. Closely related to the issue of test bias and DIF is the concept of measurement equivalence, which occurs when there are identical associations between observed test scores and latent trait across different populations

(Drasgow, 1984). It is important to evaluate tests for DIF and measurement equivalence to ensure that test findings are accurately interpreted across different samples.

Classical Test Theory and Relevant Critiques. Concepts such as the SEM and reliability estimates are associated with classical test theory (CTT), which posits a set of principles to evaluate the degree to which tests are successful at estimating unobservable variables of interests (DeVellis, 2006; Gulliksen, 1950; Lord & Novick, 1968). In CTT, an observed test score and an observed score variance are functions of a true score and an error score, as well as true score variance and error variance (Spearman, 1907, 1913).

CTT is based on a number of assumptions (Schmidt & Embretson, 2013; Zickar & Broadfoot, 2008). Briefly, the first assumption is that true scores and error scores are uncorrelated, given that errors are random. Second, a normal distribution of errors can be expected, given that errors are random and due to a combination of several factors. Therefore, the average error score is zero for each examinee in the population and across replications. Third, error scores are not correlated with scores on parallel tests or other test scores. Although CTT acknowledges measurement error, it does not generally allow for different degrees of measurement error for different ability levels (Schmidt & Embretson, 2013).

The CTT equation does not take into consideration the content or characteristics of a test item, but the theory references item relationships with other variables. Therefore, equivalent parallel forms of special test equating methods are required when a trait or construct is measured by more than one test. Strictly parallel forms have equal means, variances, and correlations with other variables. Item properties such as item difficulty and discrimination parameters have to be matched across forms. Otherwise, the true score would be dependent on particular sets of items included on a test (Schmidt & Embretson, 2013), whereby a high score would be obtained on a test with easier items, and a low score would be obtained on a test with more difficult items.

CTT has been the dominant psychometric theory over the last century. However, there are a number of limitations inherent in this theory. First, in CTT, examinee characteristics and test characteristics cannot be separated; each can only be interpreted in the context of the other. This limitation influences test precision in a number of ways. Reliability estimates vary as a function of method used and sample on which they were computed. When a test is “difficult,” an examinee will appear to have low ability whereas when the test is “easy,” the examinee will appear to have higher ability. In addition, item difficulty is defined in CTT as “the proportion of examinees who answer an item correctly” (Weiss, 1995, p.50). Therefore, item and overall test difficulty depends on the sample of examinees being measured; at the same time, examinee ability estimates depend on the degree to which the test was difficult. Consequently, a test constructed on one group of people with a given trait level cannot be used directly on a different group with a different trait level (Hambleton, Swaminathan, & Rogers, 1991; Schmidt & Embretson, 2013; Weiss, 1995).

Second, CTT computes SEM as a constant value. However, a single SEM is not an accurate reflection of a scale. Given that reliability for a given test varies depending on the group with which the individual is measured, different SEMs can be attached to an obtained score. This leads to an illogical situation whereby a person with the same score and responses is evaluated with different “precision,” depending on the group to which he/she is compared.

Third, given that CTT uses a number-correct score, test scores depend on the difficulty of items included in a test. More difficult tests result in lower average scores, and easier tests result in higher scores. Further, because difficulty levels depend on the sample, this number-correct score is dependent on the group on which the test is normed. The number-correct score artificially limits the number of levels of observed score at which a person can be assessed. For example, a test with 30 items allows only 31 possible scores. Similarly, each item in a scale is given equal weight, regardless of item difficulty, so that a correct answer to an easy item is worth as much as a correct answer to a difficult item.

Fourth, item selection procedures advocated in CTT focuses on maximizing reliability. Items are usually selected with .50 difficulty (i.e., 50% of respondents answers correctly), and further item analyses usually results in deleting items with low inter-item correlations. This results in a test with relatively equal difficulty levels and items that are highly discriminating. Such tests are effective at discriminating between upper and lower halves of population but ineffective in discriminating examinees at other levels of traits (e.g., those in the lower 10% of population).

Fifth, item parameters are regarded as fixed on a particular test in CTT. The CTT equation does not include test item characteristics or content, even though the theory underlying the CTT equation refers to item relationships with other variables. Therefore, in order to generalize a true score to other variables or tests, and to make score comparisons, it is necessary for these other variables or tests to have parallel items and it is necessary to use special test equating methods (Schmidt & Embretson, 2013).

Modern Psychometric Theory. Given the inherent limitations of CTT, it is not surprising that an alternative test development theory has been proposed. Modern psychometric theory is also known as item response theory (IRT) or latent trait theory and can be defined as a “model-based measurement in which trait level estimates depend on both the test-taker’s responses and the properties of the items that were administered” (Embretson & Reise, 2000, p.13). Fundamental to IRT is the concept of a link between item responses and the trait (known as theta, θ) measured by the scale (Drasgow & Hulin, 1990). IRT identifies item parameters like item difficulty, item discrimination, and guessing. Item difficulty, denoted by b or β, is defined as “the point along the θ continuum where individuals have a fifty percent chance of a positive response” (Drasgow & Hulin, 1990, p. 582). Item discrimination, denoted by a or α, describes how well an item can differentiate between examinees having abilities below an item location and examinees with abilities above the item location. Item discrimination is defined by the steepness of the item characteristic curve (Drasgow & Hulin, 1990). The guessing parameter, denoted by c or γ, takes into account instances where people with low θ occasionally endorse an item. The item characteristic curve graphs the relationship between changes in trait level and changes in the probability of a specified response (Cohen & Swerdlik, 2010; Hambleton, Swaminathan, & Rogers, 1991; Holland, 1990). The smaller the slope, the less discriminating the item is, because the item response probabilities (on y-axis) are relatively less responsive to changes in trait level (Embretson & Reise, 2000). Figure 1 shows item characteristic curves and the associated item parameters.

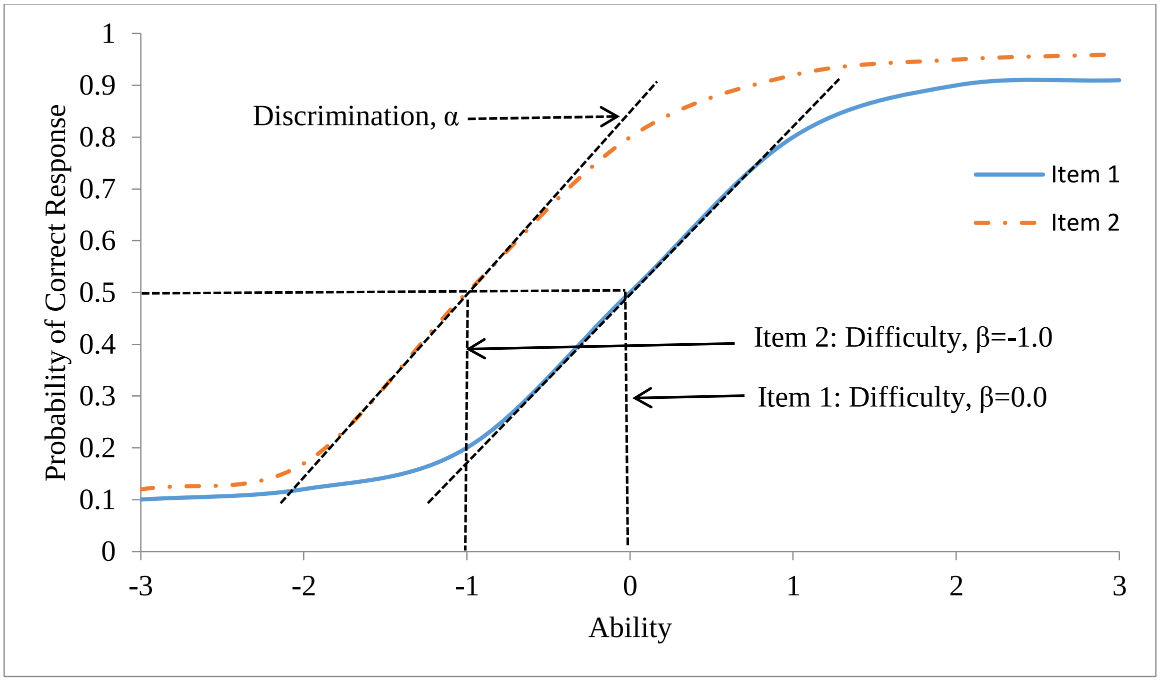


Figure 1. Item Characteristic Curves. A person with an ability level of 0.0 has a .5 chance of answering Item 1 correctly, whereas a person with a lower ability level of -1.0 has a .5 chance of Item 2 correctly. Item 1 therefore has a higher difficulty level than Item 2. The slope of the curves at the inflection point reflects item discrimination, α. Curves with steeper slopes demonstrate greater discrimination than curves with gentler slopes. The slope of Item 1 is somewhat gentler than that of Item 2, which indicates that Item 1 is less discriminating, or less responsive to changes in trait level, than Item 2.

The many benefits of IRT models provide compelling justifications for the use of such models in creating, evaluating, and applying psychological tests (Embretson & Reise, 2000). IRT incorporates techniques for evaluating the applicability of a given test across different subgroups. For example, research has shown that women tend to perform better than men on verbal memory tests (see Herlitz, Nillson, & Backman, 1997 for an overview; Lewin, Wolgers, & Herlitz, 2001; Herlitz & Rehnman, 2008). Little is known, however, regarding gender differences in how various observed memory scores discriminate levels of the underlying latent trait (memory). For example, at the same latent level of verbal memory, are there particular verbal memory tasks that yield higher scores for women than for men? IRT provides the means to more closely scrutinize test scores and their relationship to the latent variable being measured.

In IRT, differences in base rates of a given ability or trait typically do not influence results obtained and the ability estimate is not dependent upon specific items administered. IRT-derived scores reflect ability level, regardless of base rates. Consequently, there is little need for norms when using IRT-based tests, as test scores obtain meaning by comparisons of different distances from items. IRT-derived scores are also sample independent, which leads to a higher likelihood of obtaining unbiased estimates of item properties from unrepresentative samples. Rather than deriving score meaning from the location on a norm-referenced standard as is done in CTT, score meaning is derived directly from items in IRT because respondents and items are placed on a common scale. Performance can be directly inferred from the relationship between a person’s trait level and item difficulty. In addition, shorter tests can be equally as or more reliable than longer tests, when administered at appropriate θ levels. This is true for IRT-based adaptive tests, whereby different items of different levels of difficulties are administered.

A number of studies have investigated differences between scores derived through CTT and IRT. Tinsley and Dawis (1977) found that unlike CTT, IRT yielded person ability estimates that were independent of the test item difficulty levels. In other words, when students were administered two tests, their ability was estimated to be higher on the easier test and lower on the more difficult test using CTT-based raw scores. However, students’ ability estimates remained relatively constant on both the easier and more difficult tests when derived using IRT-based methods. In their Monte Carlo simulation study of test items and examinees, MacDonald and Paunonen (2002) found that IRT and CTT accurately estimated ability levels of participants (IRT-based ability parameter θ and CTT-based person test score) and test item difficulty (IRT-based *β* parameter and CTT-based item difficulty *P* value). However, the IRT-based discrimination parameter, *α*, demonstrated more consistently accurate estimates than did the CTT-based item discrimination index, particularly in simulated conditions with a large range of item difficulty statistics.

Lutz (2012) found that CTT analyses yielded significantly higher scores than did IRT on the Neuropsychological Symptom Inventory, a self-report measure of psychiatric and neurological symptoms. When considered on an individual basis, IRT yielded scores that were more closely tied to an individual's functioning level. For example, a person with a raw score of 102 always obtained a scaled score of 106 (63rd percentile) when using CTT but would obtain a score ranging from 98 to 115 (44th to 84th percentiles) depending on which items were endorsed when using IRT. The use of CTT to score items results in some missing information (e.g., item difficulty parameters associated with trait level) and runs the risk of making someone functioning at the 44th percentile look similar to someone functioning at the 63rd percentile because CTT does not take into consideration the difficulty level of the various items endorsed. In other words, sole reliance on CTT could result in a less precise interpretation of scores and a less accurate representation of a person's functioning.

Item Response Theory Applications

Given the advantages of IRT described above, successful application of IRT to various domains in psychological assessment can result in broader applicability of items across populations without requiring specific normative data for different populations (Thomas, 2011). This can result in savings of resources, especially with regards to time and the length of tests. There is also a reduction of measurement error and better evaluation of item and test bias. There can be meaningful scaling of latent variables, creation of computer adaptive tests (CATs), and more objective calibration and scaling. It is therefore not surprising that IRT has been applied to numerous domains of psychological assessment. While IRT methods have been used in clinical psychology (e.g., see Lindhiem, Kolko, and Yu, 2013; Mokros, Schilling, Eher, & Nitschke, 2012), school psychology (e.g., see Immekus & Maller, 2009; Maller, 2001; McGrew & Woodcock, 2001), and personality assessment (e.g., see Frazier, Naugle, & Haggerty, 2006; Rouse, Finger, & Butcher, 1999; Thomas & Locke, 2010; Waller, Thompson, & Wenk, 2000), this section will focus on IRT applications in neuropsychology.

IRT has been extensively applied in developing and evaluating cross-cultural and multilingual neuropsychology assessment tools (Pedraza & Mungas, 2008). For example, Gibbons et al. (2011) used various DIF-detection strategies to account for education levels and language in a 44-item naming test; they found that between four and 21 items demonstrated DIF, which impacted estimation of cognitive abilities for different cultural groups. The use of IRT enabled the researchers to identify and account for multiple sources of DIF, aspects which are typically not the focus of studies utilizing neuropsychological measures. They were also able to identify DIF-free items that can serve as test anchors when identifying item parameters and developing tests; these are important steps towards improving the validity of inferences made from increasingly common cross-cultural research and analyses. In a series of studies using somewhat similar IRT methods, Mungas and colleagues (2000, 2004, 2005) developed and refined a psychometrically sound neuropsychological test battery that matched Spanish and English language forms across 12 different scales of cognitive ability. The resulting Spanish-English Neuropsychological Assessment Scales (SENAS) did not demonstrate significant DIF in ability assessments of people across diverse demographic backgrounds. Importantly, the incorporation of IRT methods allowed for the development of a test battery with equivalent Spanish and English language forms. Such developments can facilitate unbiased measurement of cognitive ability in English- and Spanish-speaking people, and direct comparison of scores across language groups.

IRT has also been used to increase the utility of existing neuropsychological measures. One way this has been accomplished is by increasing the precision of test interpretation via decreased reliance on norms and increased precision in trait level estimations. Donnell, Belanger, & Vanderploeg (2011) demonstrated that many neuropsychological measures have scores that are not normally distributed. Reliance on norms and linear transformations for such data can lead to erroneous conclusions. IRT utilization can reduce the likelihood of error by determining latent ability rather than relying on relative performance (norms) in drawing conclusions. Precision of test interpretation can also be achieved by co-calibrating measures to estimate cognitive decline in place of standard scores. Crane and colleagues (2008) utilized IRT to co-calibrate the Mini Mental Status Exam (MMSE; Folstein, Folstein, & McHugh, 1975), the Modified Mini Mental Status Exam (3MS; Teng & Chui, 1987), the Cognitive Abilities Screening Instrument (CASI; Teng et al., 1994), and the Community Screening Instrument for Dementia (CIS “D”; Hall et al., 2000), which permitted comparisons of different cut-points of different tests used across different studies. In simulating cognitive changes over time, they found that IRT scoring was more accurate at estimating the rate of cognitive change than standard scores, particularly given that cognitive trajectories were non-linear.

Researchers have also increased the utility of neuropsychological assessment tools by using IRT as a method by which evidence for test reliability and validity can be evaluated. For example, La Femina, Senese, Grossi, and Venuti (2009) used an IRT model to demonstrate that a newly developed test battery assessing basic visuospatial abilities in children was valid and reliable. The researchers identified and eliminated misfitting items, estimated item parameters (item location on the trait continuum), and demonstrated that the newly created test fits well with a pre-specified measurement model.

IRT has also been used in the development of global measures of neurological functioning. For example, Mungas and Reed (2010) used items from nonlinear measures like the MMSE (Folstein, Folstein, & McHugh, 1975), the Blessed Information Memory Concentration Test (BIMCT; Blessed, Roth, & Tomlinson, 1968), and the Blessed-Roth Dementia Rating Scale (BRDRS; Blessed, Roth, & Tomlinson, 1968) to develop a more linear and global measure of neurological functioning. This composite measure consisted of 25 items with a more uniform distribution of item difficulty across a wide range of ability. Similarly, Mungas, Reed, and Kramer (2003) utilized IRT to obtain a global cognition score from the Word List Learning Test of the Memory Assessment Scale (MAS; Williams, 1991), WMS-R (Weschler, 1987), Digit Span Forward and Backward, animal category fluency (Morris et al., 1989; Welsh et al., 1994), and letter fluency (FAS

Test; Benton & Hamsher, 1976) scores. Compared to the Mattis Dementia Rating Scale (Mattis, 1988), which had high reliability at low ability levels but low reliability for more normal cognitive functioning, the global cognition score had consistently high reliability levels over a wide range of ability levels.

IRT can also improve neuropsychological evaluations by reducing test length (Calamia, Markon, Denburg, & Tranel, 2011). Given the emphasis of modern day managed health care on quick and effective services, this could be one of the more significant contributions of IRT to neuropsychology. Schultz-Larsen, Kreiner, & Lomholt (2007a, 2007b) used an IRT model to evaluate the MMSE (Folstein, Folstein, & McHugh, 1975) and to derive a shorter, nine-item MMSE measure that has similar sensitivity, specificity, and predictive values as the original, 11-item measure. Calamia and colleagues (2011) used IRT to estimate the difficulty level for each item on the Judgment of Line Orientation test (Benton, Sivan, Hamsher, Varney, & Spreen, 1994), a measure of visual spatial reasoning. The use of IRT aided in item selection for shorter administration while maintaining comparability to the full form. Specifically, the researchers reduced test length by almost 10 items (a third of the 30-item test) by incorporating a start-point at item 19, and establishing basal and ceiling rules set at 6 items. Pearson correlation between the shorter and full form was .96. Previous versions of short-forms led to differences in impairment classification of between four to 11 percent compared to the full form (Qualls, Bliwise, & Stringer, 2000; Winegarden et al., 1998); this newer administration method led to differences in impairment classification of three percent, compared to full form administration. Spencer and colleagues (2013) found this

IRT-derived short form to be superior to other short forms. Specifically, this version had adequate reliability and similar diagnostic classification statistics as the 30-item test.

Item response theory in verbal memory tasks. The prior sections highlighted the many benefits of IRT and various ways in which it has been applied. Researchers have called for further application of IRT in clinical psychology assessments in general (e.g., Embretson & Reise, 2000; Rouse, Finger, & Butcher, 1999), and neuropsychology (e.g., Gavett & Horwitz, 2011; Mungas & Reed, 2010) in particular. Given the widespread use of verbal learning measures in the assessment of dementia, it is apt to consider how IRT may be used to evaluate and enhance such measures. It is well-known that not all items function equally on a memory test. For example, the tendency for people to remember the first few (primacy effect) and last few words (recency effect; Deese & Kaufman, 1957) on a list-learning task may make some items "easier" than others. Similarly, emotional salience in a story memory test could make some items more easily recalled than others (Kensinger & Corkin, 2003). This study will focus on two commonly used measures of verbal memory, the CVLT-II, and the LM subtests of the WMS-IV.

The California Verbal Learning Test – Second Edition (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000). The CVLT assesses immediate and delayed recall of a 16-item word list. It is among the most widely used measures among neuropsychologists, both historically and presently (Lees-Haley, Smith, Williams, & Dunn, 1996; Rabin, Barr, & Burton, 2005). The CVLT-II measures recall and recognition of two word lists over a number of immediate- and delayed-memory trials (Delis, Kramer, Kaplan, & Ober, 2000). The CVLT-II provides various measures to assess different aspects of learning and memory. One such measure is primacy and recency recall. This pertains to a common phenomenon in free recall learning tests, known as the serial position effect (Deese & Kaufman, 1957), whereby the probability of a word being recalled is influenced by its position on the list. Specifically, during free recall of a word list, the middle items are least frequently recalled, the first items are moderately well recalled, and the last items are most frequently recalled. Atkinson and Shiffrin (1965) posited in their buffer model of memory that the primacy effect is facilitated by the increased opportunity for items at the beginning of the list to be rehearsed and encoded into long term memory stores. Given that the buffer is a constant size and contains only a limited number of items, earlier-presented items leave the buffer (either lost, forgotten, or entered into long term memory) as new items are presented. They proposed that the recency effect occurs because items at the end of the list are still present in the rehearsal buffer at the time of recall. The middle items do not have as much time to be rehearsed and encoded into long term memory stores, and are not sufficiently recent to still be present in the rehearsal buffer at the time of recall.

Some researchers have argued that list learning tasks that are scored based on total number of words recalled, without taking into consideration the serial position effect, lack construct validity to some degree (Buschke et al., 2006; Gavett & Horwitz, 2011). This is because such a scoring approach does not take into consideration that clinical manifestations of Alzheimer’s disease demonstrate a disruption in episodic memory (Sperling et al., 2010) but a general sparing of verbal attention (Linn et al., 1995), which translates into a reduction in the primacy effect but preservation of the recency effect (Bigler, Rosa, Schultz, Hall, & Harris, 1989). Recognizing this, Buschke and colleagues (2006) assigned a weighted scoring system so that words recalled from short-term memory (primacy effects) were scored with greater weights than those recalled from basic attention (recency effects). Their scoring system resulted in better discrimination of mild Alzheimer's disease from controls, compared to an unweighted scoring system.

Gavett and Horwitz (2011) used IRT to determine the degree to which the serial position effect posed a threat to the construct validity of the Rey Auditory Verbal Learning Test (Rey, 1964), a list learning measure of verbal episodic memory that is conceptually similar to the CVLT. They used item-level data for a single immediate recall trial and found that four of the 15 items did not fit their model, as indicated by an approximate chi-square distribution. They identified and eliminated items that appeared to be more related to attention (as opposed to memory) and found that such procedures led to a better fit to their data and reduced the confounding effects of attention on the test. In other words, removing these items reduced the degree to which the primacy and recency effects violated assumptions of the IRT model (unidimensionality and local independence), which led to a more unidimensional model of episodic verbal memory, and more accurate parameter estimations. Given these findings and the widespread use of the CVLT, it would be beneficial to determine if similar adjustments (weighted scoring system, exclusion of items that misfit the IRT model) to the CVLT-II may be warranted.

The Logical Memory (LM) subtest of the Wechsler Memory Scales – Fourth Edition (Wechsler, 2009). The WMS-IV LM subtest is another frequently administered measure of verbal memory (Rabin, Barr, & Burton, 2005), which involves the immediate and delayed recall of two short stories, along with a yes/no recognition trial for the two stories following delayed recall.

Similar to list learning tasks, the serial position effect has been observed in contextual memory tasks, such as paragraph recall (Hall & Bornstein, 1991; Messier, Gagnon, & Knott, 1997). In such tasks, studies have found that both normal controls and patients with closed-head injury demonstrate primacy and recency effects. However, normal controls better recall the middle portion of the story/paragraph, and demonstrate less prominent primacy and recency effects. Unlike a list-learning test, however, people do not tend to recall the last words of a passage prose first, and the middle words last (Deese & Kaufman, 1957). Thus, although people may recall more information from the beginning and ending portions of a paragraph than the middle, they tend to recall information in logical order, going from beginning to middle, to end. To date, no known studies have attempted to determine an optimal scoring approach or norms that incorporate primacy and recency effects in the LM test. It is unclear if a weighted scoring system or the exclusion of some items in the LM test may improve test precision.

The Present Study

The proposed study builds upon these research findings on the applicability of IRT models to various psychological assessment measures. Specifically, this study will utilize IRT models to identify item properties of the CVLT-II and the WMS-IV LM subtest. Three primary research questions will be addressed:

(1) What are the item difficulty and discrimination parameters for CVLT-II and LM test items across different recall trials (immediate and delayed recall)? Given the serial position effect, it is hypothesized that items from the middle

portion of the word list (CVLT-II) and stories (LM) will have greater item difficulty and discrimination values than items from the first and last portions of the word list and stories.

(2) How do the item parameters for the CVLT-II compare to those of the LM subtests? Is one test more effective in quantifying memory functioning? In other words, is there an advantage for a test because it more optimally evaluates across the spectrum of memory ability? Few studies have evaluated the effectiveness of CVLT-II tests compared to LM subtests in quantifying memory. However, in general, studies have found that list-learning tests performed better than story recall tests in assessing and discriminating between different levels of cognitive functioning (see Helmstaedter, Wietzke, and Lutz, 2009; Rabin et al., 2009; De Jager et al., 2003). Therefore, it is hypothesized that CVLT-II will be more effective in quantifying memory functioning (i.e., higher discrimination values across the range of memory abilities) compared with LM.

(3) What, if any, weighted item scoring approach will lead to optimal scoring and quantification of memory functioning? For example, Buschke and colleagues (2006) assigned greater weights to items from the first and middle portions of the list/stories and less weight to items at the end of the list. It is hypothesized that assigning greater weight to items with higher difficulty parameters will improve discrimination of cognitive decline (quantified by discriminating patients in a clinical setting from participants recruited in a research setting),

compared to an unweighted scoring approach. Improved discrimination will be quantified using receiver operating characteristics (ROC) curves.

METHODS

Participants

Archival data from 776 individuals were obtained for this study. Of these, 526 (67.8%) were research data obtained from college students and 250 (32.2%) were clinical data obtained from a neuropsychology assessment clinic. College students completed the CVLT-II and/or WMS-IV LM subtests as part of an extended neuropsychology battery investigating measures of verbal and nonverbal memory. These data were collected from 2011 to 2014. College students had a mean age of 19.22 (SD=1.34), education level of 12.8 years (SD=1.10), and mean Wechsler Test of Adult Reading (WTAR; Wechsler, 2001) standard score of 110.37 (SD=10.63). The majority of college students in this study were Caucasian (72.6%). A minority of participants endorsed receiving a psychiatric (3.4%), learning disorder (1.9%), and/or Attention-Deficit/Hyperactivity Disorder (ADHD; 2.8%) diagnosis. Patients at a neuropsychology assessment clinic presented with various cognitive difficulties, and were seen in the clinic from 2008 to 2014. Patients were referred to the clinic to assess cognitive functioning relating to memory complaints, multiple sclerosis, traumatic brain injury, difficulties with attention, mood disorders, and seizures, among others. Patients from the neuropsychology assessment clinic had a mean age of 39.69 (SD=13.39), education level of 14.00 (SD=2.78), and mean Wide Range Achievement Test – Fourth Edition Word Reading subtest (WRAT-4; Wilkinson & Robertson, 2006) standard score of 97.46 (SD=12.99).

Inclusion criteria are the completion of the CVLT-II and/or WMS-IV LM. Given the archival nature of these studies, not every research participant, or every patient at the neuropsychology clinic were administered both the CVLT-II and WMS-IV LM. The decision to administer select tests is typically made a priori, based on the research question or presenting problem. Of the 776 individuals initially included in this study, 428 (55.2%) completed both the CVLT-II and LM subtests, 170 (21.9%) completed only the CVLT-II, and 178 (22.9%) completed only the LM subtests.

Individuals whose scores on various effort measures indicated suboptimal effort were excluded from further analyses. A Trial 2 score of less than 45 (out of 50) on the Test of Memory Malingering (TOMM; Tombaugh, 1996; Bianchini, Mathias, & Greve, 2001; Moore & Donders, 2004), a score of less than 16 (out of 24) on either the easy or hard trials on the Victoria Symptom Validity Test (VSVT; Slick, Hoop, Strauss, & Thompson, 1997; Slick, Hopp, Strauss, & Spellacy, 1996), or a total score of less than 15 (out of 16) on the CVLT-II forced recognition trial could indicate suboptimal effort (Jacobs & Donders, 2007; Moore & Donders, 2004). Of the 776 individuals, 66 (8.51%) were not administered any symptom validity test, and 21 (2.71%) demonstrated some indication of suboptimal effort. These 21 individuals were therefore excluded from further analyses; these individuals did not differ in age, education level, or word reading achievement compared to the overall sample. Analyses were conducted on a sample of 755 individuals, of which 516 (68.3%) were college students, and 239 (31.7%) were patients at a neuropsychology clinic. Data analyzed were based on 577 CVLT-II protocols and 589 LM protocols. Given that IRT parameters are not sample-dependent and IRT analyses require responses across the entire spectrum of ability levels, all IRT analyses were based on the entire sample obtained (no separate IRT analyses were conducted to compare between the research and clinical samples).

Sample size justification. There is no definitive rule that specifies what sample size is needed to conduct various IRT models. However, more complex IRT models require larger sample sizes. Sample sizes used in IRT studies range from 105 (Mokros et al., 2012) to 32,000 (Reise & Waller, 2003). A two parameter logistic (2 PL) model (described in detail in the Procedures section) is selected because it provides respective item discrimination and difficulty parameters. The current sample sizes (577 and 589) are adequate for the complexity of this 2 PL model based on a Monte Carlo simulation study by Hulin, Lissak, and Drasgow (1982). Specifically, Hulin and colleagues simulated binary item responses (1 = correct, 0 = incorrect) for various sample sizes (200, 500, 1000, and 2000) and test lengths (15, 30, and 60 items) and compared the root mean squared error (RMSE) between recovered and actual item characteristic curves for various IRT models (1-, 2-, and 3-parameter logistic models). Their findings suggested that a sample size of 500 was adequate for a 2 PL model, with RMSE values of .07 for the 15-item test, and .04 for the 30-item test. The suitability of this 2 PL model was further evaluated using Yen's (1981) Q_1 statistic and Xcalibre standardized residual *z* statistic (see Statistical Analyses section below).

Procedures

Item-level responses for both the CVLT-II and the LM subtests were obtained. Data was coded in a binary fashion to indicate whether each item in each trial was correctly recalled by the participant (0 = incorrectly or not recalled; 1=correctly recalled). For the CVLT-II, responses on the immediate recall for Trials 1 and 5, short delay free recall (SDFR), and long-delay free recall (LDFR) were used for this study. For the LM subtests, participants' responses to each of the 25 items in the immediate and delayed recall trials of stories B and C were used for this study.

Measures

CVLT-II. As previously described, the CVLT-II measures recall and recognition of two word lists over a number of immediate- and delayed-memory trials (Delis, Kramer, Kaplan, & Ober, 2000). The examinee is asked to recall words from List A immediately after each presentation of the list in the first five learning trials. List A consists of 16 words, with four words from each of the four semantic categories of furniture, vegetables, animals, and ways of transportation. An interference list (List B) is then presented for one trial. List B also consists of 16 words, with four words from each of the four semantic categories of musical instruments, animals, vegetables, and areas of the house. The interference list is then followed by SDFR and short-delay cued-recall trials of List A. After a 20-minute delay, the LDFR, long-delay cued-recall, and yes/no recognition trials of List A are administered. The CVLT-II also includes an optional forced-choice recognition trail, which may be administered after another 10-minute delay.

Numerous studies have demonstrated the construct validity of the CVLT as a measure of episodic verbal learning and memory (see Alexander, Stuss, and Fansabedian, 2003; Baldo, Delis, Kramer and Shimamura, 2002; Bondi et al., 1994; Chen, Kareken, Fastenau, Trexler, & Hutchins, 2003; Crosson, Novack, Trenerry, & Craig, 1988; Gold et al., 1994; Massman et al., 1992). Factor analyses suggests that the CVLT-II evaluates (1) Attention Span, which includes measures from List A Trial 1, and List B, (2) Learning Efficiency, which includes measures from List A Trial 5, semantic clustering, and recall consistency, (3) Delayed Memory, which includes measures from short- and long-delay free- and cued-recall, and recognition hits, and (4) Inaccurate Memory, which include measures from total intrusions and recognition false positives (Donders, 2008; DeJong & Donders, 2009). The CVLT-II demonstrates good psychometric properties (Delis, Kramer, Kaplan, & Ober, 2000). One-month test-retest reliability for the standard-standard forms and the standard-alternative forms were within the range of *r* = .80 to .84, and *r* = .61 to .73 respectively. Practice effects for the standard-standard forms and the standard-alternative forms were within Cohen's *d* ranges of 0.27 to 0.61 and -0.01 to 0.18 respectively (Woods, Delis, Scott, Kramer, & Holdnack, 2006).

WMS-IV LM. As previously described, the WMS-IV LM subtests involve the immediate and delayed recall of two short stories, along with a yes/no recognition trial for the two stories following delayed recall. This test has been used to assess for Alzheimer's disease (Brown & Storandt, 2000; Cullum, Butters, Troster, & Salmon, 1990), brain injury (Bigler et al., 1996; Kaitaro, Koskinen, & Kaipio, 1995), mild cognitive impairment (Nordlund et al., 2007, 2008), and temporal lobe epilepsy (Kent et al., 2006), among others. In general, clinical samples perform worse than control groups, and in the impaired range on this measure (Strauss, Sherman, & Spreen, 2006). With respect to construct validity, a factor analytic study by Hoelzle, Nelson, and Smith (2011) demonstrated that the WMS-IV has an invariant two-dimensional structure that reflects auditory memory and learning (consisting of the LM and VPA subtests), and visual attention and memory (consisting of the Designs, Visual Reproduction, Spatial Addition, and Symbol Span subtests). The WMS-IV LM demonstrates sound psychometric properties, with reliability estimates ranging from .77 to .96 (Wechsler, 2009).

Convergent validity is moderate to high between the LM and other verbal memory measures like the WMS-IV Verbal Paired Associates (VPA) subtests and the CVLT-II (Wechsler, 2009).

Data Analyses

Data analyses were conducted using Xcalibre 4 (Assessment Systems Incorporated; Guyer & Thompson, 2012) for IRT analyses, R version 3.1.1. (R Development Core Team, 2014), including the ltm package (Rizopoulous, 2006) for modified parallel analyses and PEIP package (Lees, 2014) for Q_1 analyses, and MedCalc for Windows, version 14.12.0 (MedCalc Software, Ostend, Belgium, 2014) for comparative analyses of ROC curves. SPSS Version 21 (IBM Corp, 2012) was used for all other analyses. For all significance tests, alpha level was set at .05.

To investigate the serial position effect, mean scores for all items in each trial (16 items in CVLT-II trials and 25 items in LM trials) were regressed onto the item sequence number (1 to 16 or 1 to 25) using a quadratic term, consistent with the methodology utilized by Gavett & Horwitz (2011).

Numerous IRT models exist. Key factors that will influence one’s choice of IRT models include the scale of response data and dimensionality. Scale of response data could be binary, polychotomous, or continuous. Given that responses on the CVLT-II and LM are scored in a binary fashion (correct or wrong), a logistical IRT model was utilized. Another factor which informs the selection on an IRT model is dimensionality of the item pool or latent traits underlying item responses (Drasgow & Hulin, 1990). Unidimensional models are more commonly used and assume that items assess only one latent trait, whereas multidimensional models assume that items assess more than one latent trait.

Given that both the CVLT-II and LM are measures of verbal memory (as opposed to a battery measuring verbal, visual, working memory), a unidimensional IRT model was utilized.

Three types of unidimensional, logistical IRT models exists. The one-parameter logistic model (1 PL) provides only item difficulty estimates. The two-parameter logistic model (2 PL) provides item difficulty and discrimination estimates. The three-parameter logistic model (3 PL) provides an additional estimate for the guessing parameter, which is the likelihood that someone with a low ability would be able to answer an item correctly simply from guessing. Given that this study hopes to determine not only item difficulty, but also the degree to which items of different trials and tests discriminate individuals of varying verbal memory abilities, a 2 PL model was used to estimate the difficulty and discrimination parameters of each item. Guessing was not a significant concern, given that questions were open ended in nature. The 2 PL model can be represented by the formula:

$$\mathrm{P}(X_{is} = 1 \mid \theta_{\mathrm{s}}, \beta_i\,, \alpha_i) = \frac{\exp[\alpha_i(\theta_{\mathrm{s}} - \beta_i)]}{1 + \exp[\alpha_i(\theta_{\mathrm{s}} - \beta_i)]}$$

The 2 PL IRT item difficulty and discrimination estimates were used to plot the item characteristic curves, item information curves, and test information curves. As described above, an item characteristic curve graphs the relationship between changes in trait level and changes in the probability of a specified response. The item information curve indicates the quality of a given item, based on how closely the difficulty of the item matches the ability of the respondent. Finally, the test information curve indicates the quality of a given test (in this study, each of the 8 trials investigated), and represents the sum of item information curve values for each trait level. All IRT analyses were focused on estimating ability levels (θ) ranging from 3 SD above to 3 SD below the mean of an assumed normal distribution (M = 0.00, SD = 1.00).

IRT has a number of basic assumptions (Cohen & Swerdlik, 2010). First, unidimensionality is the assumption that the scale measures a single continuous latent construct. Although this assumption cannot be strictly met (motivation, test anxiety, guessing are among factors that influence test performance), in IRT, this assumption is considered to be adequately met if a set of test data demonstrates the presence of a "dominant" component or factor that influences test performance (Hambleton, Swaminathan, & Rogers, 1991). The unidimensionality assumption was tested using modified parallel analysis. Parallel analysis references the distributions of eigenvalues from correlation matrices arising from random data. It enables the comparison of eigenvalues from an empirical correlation matrix with the value of the distribution of random eigenvalues, which allows the filtering of meaningful factors from random variations (Tran & Formann, 2009; Weng & Cheng, 2005). Parallel analysis assumes that observed variables are linear functions of the factors (Humphreys & Montanelli, 1975). Given that dichotomously scored item responses do not satisfy this assumption, modified parallel analysis is used. Modified parallel analysis differs from traditional parallel analysis in three aspects: (1) tetrachoric correlations are used, as opposed to product moment correlations, (2) the largest off-diagonal correlations are taken as communality estimates, as opposed to squared multiple correlations, and (3) dichotomously scored synthetic variables that satisfy the unidimensionality assumption of IRT are used, as opposed to continuous synthetic variables that are statistically independent. Modified parallel analysis employs Monte Carlo simulation to compare the matrix of tetrachoric correlations obtained from the actual data under a specified IRT model to the simulated data (averaged across 100 simulations). Resultant second eigenvalues from real and randomly-generated data were compared, with the null hypothesis that there will be no meaningful difference between the second eigenvalues of the observed data and the simulated data. Resultant screeplots were also visually examined and compared to the exemplars in Dragsow and Lissak (1983) to determine if the IRT models used were robust against potential violations of the unidimensionality assumption.

Second, local independence is the assumption that there is a systematic relationship between all test items, and that the relationship pertains to a given level of construct. When the latent trait influencing test performance is held constant, an examinee's responses to any pair of items are statistically independent. Considered together with the unidimensionality assumption, this means that the complete latent space consists of one construct or trait. Item fit was examined using the Q_1 statistic (Yen, 1981), which approximates the chi-square (χ^2) distribution. Specifically, the procedure to conduct a Q_1 analysis involves (1) estimating θ (in this study, memory ability) and item parameters from a dataset, (2) sorting examinees by their θ estimates, (3) forming subgroups of the sorted examinees, (4) calculating the proportion of examinees in each subgroup who answered correctly/incorrectly for each item, and (5) comparing these "observed" proportions with those predicted by the model using a χ^2-like statistic or graphical representation (Ankenmann, 1994). However, due to the very large size of the sample and associated power, it is likely that most items fit statistics will be found to over- or under-fit the model to a statistically significant degree (Reise, 1990). Therefore, the standardized residual *z* statistic provided by Xcalibre software, which is more robust against effects of sample size, was also used to evaluate item fit. Items that were found to misfit the model using both Q_1 and standardized residual *z* statistics will be excluded from IRT analyses. To test local independence, correlations among item position, item discrimination, and squared item discrimination were calculated using item parameters generated by 2 PL IRT analyses. Strong correlations between item discrimination and item position indicate that the assumption of local independence has been violated, given that it indicates that the relationship between test items are not only influenced by ability levels, but also item position on the list (Reise & Waller, 1990).

Third, monotonicity is the assumption that the probability of endorsing an item (indicative of higher levels of a particular construct or trait) increases as the underlying level of construct increases. This monotonicity can be characterized with an item characteristic curve, which graphs the relationship between changes in trait level and changes in the probability of a specified response (Cohen & Swerlik, 2010; Hambleton, Swaminathan, & Rogers, 1991; Holland, 1990). Item characteristic curves were plotted for each item to examine monotonicity.

Finally, to investigate if a weighted scoring approach could increase test precision, items were assigned different scoring weights based on item difficulty level identified from IRT analyses. Utilizing Gavett and Horwitz’s (2011) suggestion to assign scores based on item difficulty levels, items with the lowest difficulty level received the lowest score (1 point) and items with the highest difficulty level received the highest score (16 for CVLT-II items, 25 for LM items). Receiver operating characteristics (ROC) analyses were conducted to compare test discrimination between these two scoring methods in distinguishing healthy, high-functioning young adults (research sample) from a clinical sample of people presenting with cognitive complaints at a neuropsychology clinic. ROC graphs are two-dimensional graphs in which the true positive rate is plotted on the Y axis and the false positive rate is plotted on the X axis (Fawcett, 2006). An ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives). The area under the curve (AUC) represents test accuracy. AUC values range from 0.5, corresponding to very low accuracy (pure chance), to 1.0, corresponding to perfect accuracy (Faraggi & Reiser, 2002; Grzybowski & Younger, 1997). Swets (1988) provided guidelines for interpreting the AUC, whereby an AUC value in the range of 0.5 to 0.7 represents no to low discriminatory power, an AUC value in the range of 0.7 to 0.9 represents moderate discriminatory power, and an AUC value >0.9 represents high discriminatory powers.

RESULTS

CVLT-II and LM means and standard deviations are presented in Table 1. As can be seen, the clinical sample consistently scored lower than the research sample, across all verbal memory trials investigated. Each test trial was evaluated to examine serial position effect (quadratic regression), unidimensionality (modified parallel analysis), item parameters (2 PL IRT analyses), item fit (Q_1 and standardized residual *z* statistics tests), monotonicity (item characteristic curves), local independence (item-position correlations and squared correlations), and differences in test precision utilizing a weighted scoring approach (ROC analyses). Results will be presented sequentially by test and trials within tests.

Table 1 Mean, standard deviations, t-test results, and effect sizes between research and clinical samples' standard scores

Verbal Memory Scores	Research Participants Means (SD)	Clinical Sample Means (SD)	*t*	Cohen's *d*
CVLT-II Trial 1 z-score	-0.13 (1.43)	-0.56 (1.13)	3.89*	0.34
CVLT-II Trial 5 z-score	0.36 (2.10)	-0.25 (1.50)	3.87*	0.34
CVLT-II Trial SDFR z-score	0.29 (0.97)	-0.27 (1.16)	6.09*	0.52
CVLT-II Trial LDFR z-score	0.19 (0.90)	-0.34 (1.22)	5.73*	0.50
LM I standard score	10.39 (2.71)	9.54 (3.30)	3.28*	0.28
LM II standard score	10.17 (2.64)	9.19 (3.44)	3.69*	0.32

Note. z-scores have a mean of 0 and SD of 1; Standard scores have a mean of 10 and SD of 3.
CVLT-II = California Verbal Learning Test; SDFR = Short-Delay Free Recall; LDFR = Long-Delay Free Recall; LM = Logical Memory.
* $p < .05$.

CVLT-II Trial 1

To investigate the serial position effect, mean scores for all 16 CVLT-II Trial 1 items were regressed onto the item sequence number (1 to 16). As indicated in Figure 2, the serial position effect was apparent on CVLT-II Trial 1 (F = 28.01, *p* < .01) utilizing a quadratic term (nonlinear regression). Items presented at the beginning and end of the word list were recalled at a higher frequency than words presented in the middle of the list.

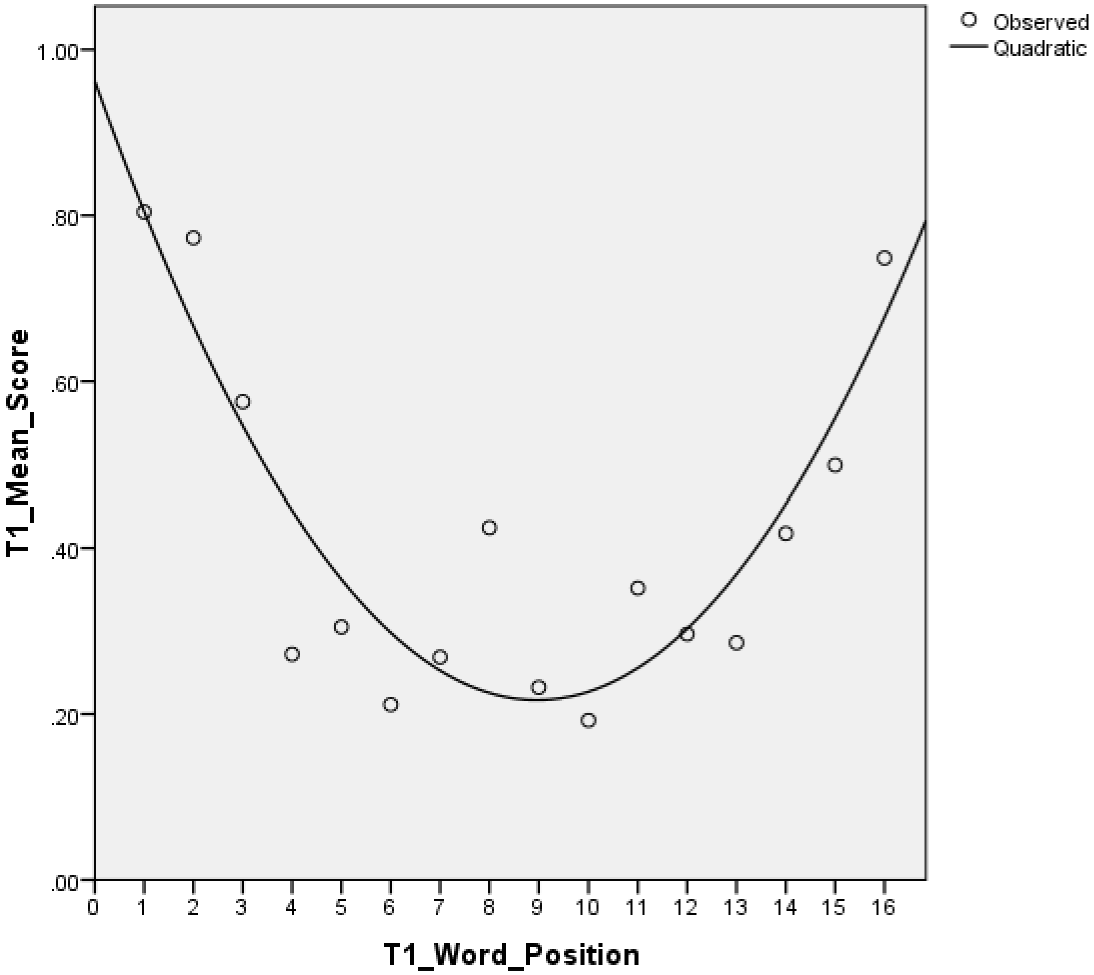


Figure 2 Regression model for CVLT-II Trial 1

As described above, it is important to evaluate if assumptions underlying the IRT model are met, prior to conducting IRT analyses. Unidimensionality is the assumption that the scale measures a single continuous latent construct and is tested using modified parallel analysis. For unidimensionality to be assumed, the first eigenvalue from the real data should be much larger than the first eigenvalue from the randomly-generated data, whereas subsequent eigenvalues from the real and randomly-generated data should be similar. The CVLT-II Trial 1 scree plot was also visually examined to evaluate whether a strong first factor is present.

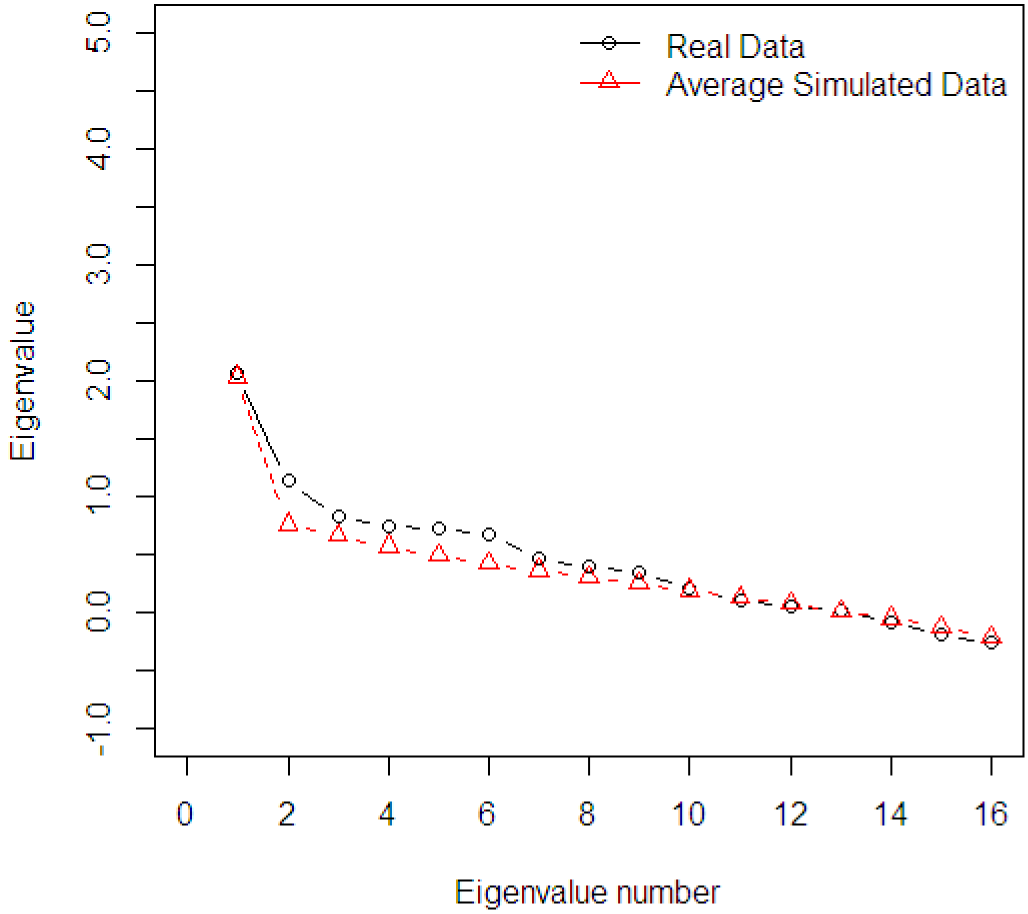


Figure 3 Scree plot for modified parallel analysis of CVLT-II Trial 1.

As can be observed in Figure 3, there was not clear support for the unidimensionality of CVLT-II Trial 1. When compared with the simulated data, the second eigenvalue of the observed data (1.13) was significantly larger than the second eigenvalue of the simulated data (0.76, $p < .01$). It is not uncommon for researchers to identify that modified parallel analysis does not technically support a unidimensional model (see Childs et al., 2000; Gavett & Horwitz, 2011; Thomas & Locke, 2010 for examples). Next, it is routine to evaluate the severity of the violation. Despite evidence that CVLT-II Trial 1 may have multiple underlying dimensions, the violation of the assumption of unidimensionality is not so severe as to invalidate the use of a unidimensional IRT model according to visual guidelines set forth by Drasgow and Lissak (1983, p.364), who acknowledged that the assumption of unidimensionality is "too strong for most data sets from applied psychology." Therefore, 2 PL IRT analyses were conducted.

The proportion of examinees that answered each item correct (P) is presented in Table 2. Findings from the 2 PL IRT analysis that provide estimates of item difficulty (β) and discrimination (α), as well as item fit statistics (Q_1 and z) are also presented in Table 2. The focus of the IRT analysis was on estimating ability levels (θ) ranging from 3 SD below to 3 SD above the mean of an assumed normal distribution with a mean of 0 and a standard deviation of 1. The Q_1 statistic approximates the χ^2 distribution (Yen, 1981), and was used to investigate item fit (the degree to which observed/obtained data matches predicted data; Orlando & Thissen, 2000).

Table 2 IRT item parameters, Q_1, and *z* item fit statistics for CVLT-II Trial 1

Item	CVLT-II Trial 1					
	P	β	α	Q_1	*p*	*z*
1	0.80	-1.82	0.46	12.31	<.01*	0.63
2	0.77	-1.44	0.52	9.45	.02*	0.54
3	0.58	-0.28	0.69	22.28	<.01*	0.40
4	0.27	1.76	0.33	3.97	.26	0.64
5	0.31	1.25	0.41	4.24	.24	0.23
6	0.21	2.05	0.39	3.61	.31	0.61
7	0.27	1.57	0.39	2.45	.48	0.34
8	0.43	0.55	0.35	9.47	.02*	0.35
9	0.23	1.92	0.37	4.48	.21	0.60
10	0.19	2.10	0.41	6.41	.09	0.56
11	0.35	1.12	0.33	3.41	.33	0.49
12	0.30	1.50	0.35	2.85	.42	0.55
13	0.29	1.53	0.36	7.58	.06	0.56
14	0.42	0.87	0.23	16.33	<.01*	1.13
15	0.50	0.07	0.22	32.64	<.01*	1.07
16	0.75	-2.02	0.30	33.89	<.01*	1.29

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
**p* < .05.

Based on the Q_1 statistic, seven of 16 items in CVLT-II Trial 1 did not fit the IRT model (items 1, 2, 3, 8, 14, 15, and 16). The misfit between item statistics and the model is likely due to the large sample size and associated statistical power (Reise, 1990; Gavett & Horwitz, 2011). Given this finding, and as recommended by the Xcalibre software manual, an alternative fit statistic, the standardized residual (*z*) fit statistic, can be also used to evaluate the significance of item misfit, as it is less sensitive to the effects of sample size (Guyer & Thompson, 2012, p.46). All 16 items appeared to fit the model, according to the standardized residual (*z*) statistic.

As can be seen in Table 2, items in the middle of the list had higher difficulty (β) parameters than items at the beginning and end of the list, which is expected given primacy and recency effects. Items at the beginning of the list had higher discrimination levels than items in the middle and items at the end of the list. Item characteristic curves are presented in Figure 4, whereby each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination. Item 16 (cabbage) was the easiest item, whereby even an examinee with a very low ability level ($\theta = -2.02$) has a 0.5 likelihood of answering that item correctly. In contrast, item 10 (lamp) had the highest difficulty level in Trial 1, whereby even an examinee with an above-average ability level of $\theta = 2.10$ has only a 0.5 likelihood of answering that item correctly. Item 3 (giraffe) has the steepest slope, indicating that it most sharply discriminates between examinees, compared to other items, and that it functions best when discriminating between examinees with ability levels ranging around $\theta = -0.28$. Items earlier on the list had higher discrimination levels (first four words mean $\alpha = 0.50$) than items in the middle (middle eight words mean $\alpha = 0.38$) or end (last four words mean $\alpha = 0.28$) of the list, indicating they most effectively discriminate between different levels of memory ability in Trial 1. As the item characteristic curve indicates, monotonicity (probability of answering an item correctly increases as trait level increases) was evident across all 16 items of CVLT-II Trial 1.

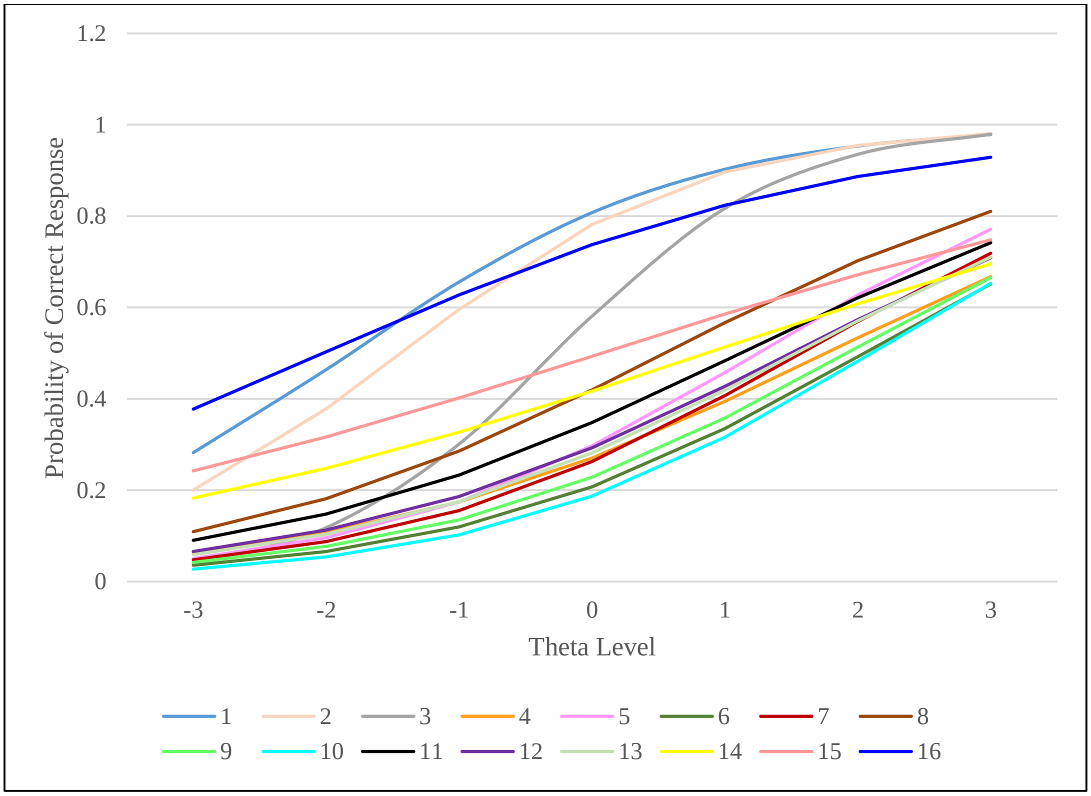


Figure 4 Item characteristic curve for CVLT-II Trial 1.

Item information curves are presented in Figure 5, whereby each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent. Items 14, 15, and 16 have low utility in terms of providing information about examinees, given the relatively flat curve across the ability continuum.

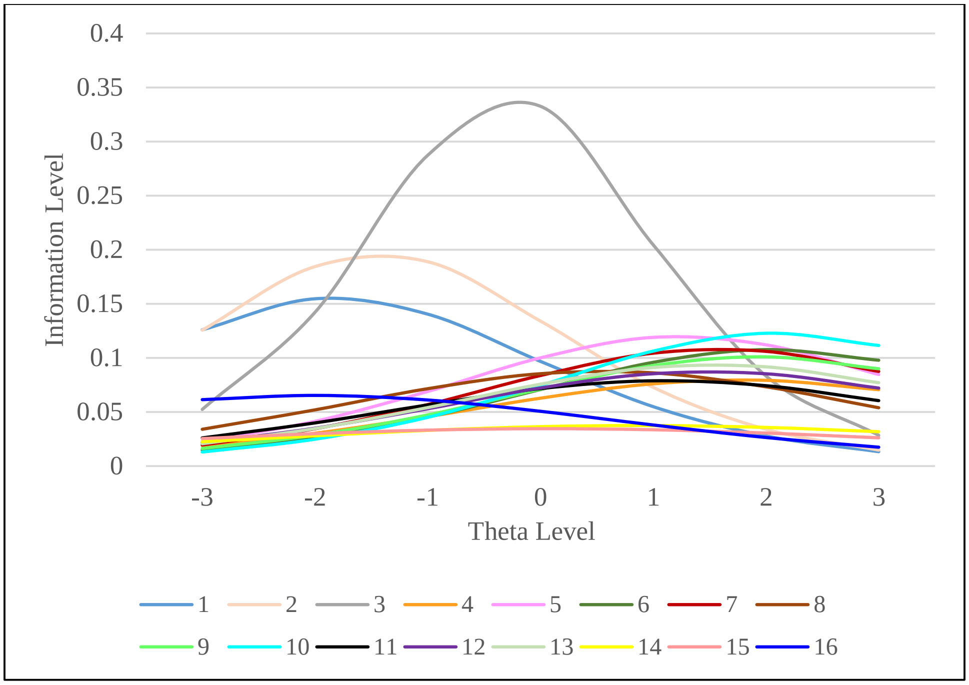


Figure 5 Item information curves for CVLT-II Trial 1.

The test information curve (which provides information about the quality of a given test or trial and represents the sum of item information curve values for each trait level) is presented in Figure 6. The curve indicates that the maximum information that can be drawn from CVLT-II Trial 1 is approximately 1.45, at a latent memory ability level of 0.10.

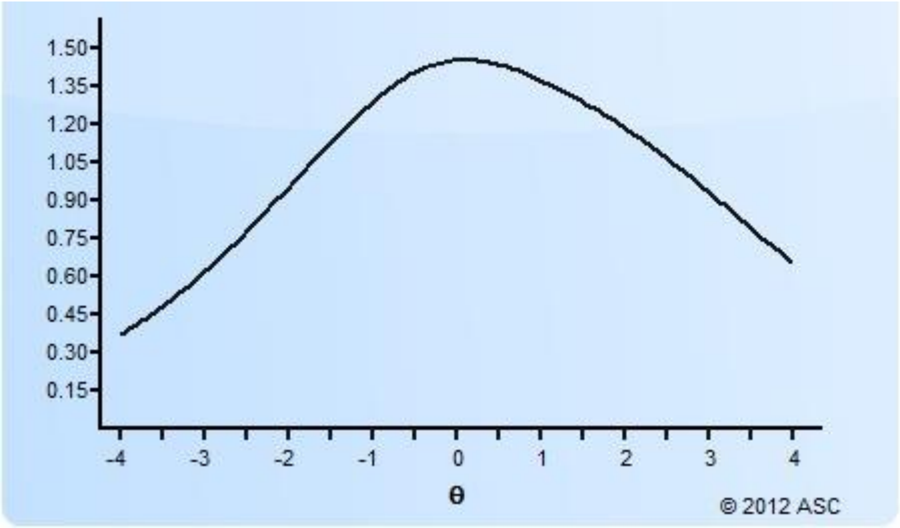


Figure 6 Test information curve for CVLT-II Trial 1.

Local independence is the assumption that there is a systematic relationship between all test items, and that the relationship pertains to a given level of construct (Hambleton, Swaminathan, & Rogers, 1991). To check if local independence is assumed, correlations among item position, item discrimination, and squared item discrimination were calculated, given there are likely to be strong associations between item discrimination and item position (Reise & Waller, 1990). The parameter estimates for the 16-item version of CVLT-II Trial 1 are likely to be somewhat biased by a violation of the local independence assumption, given that correlations between item position and item discrimination were quite high (r = -.74, p < .01; r^2 = -.67, p = .01). Given this finding, item parameter estimates may be slightly biased; specifically, the degree to which an item discriminates between different ability levels are not solely dependent on verbal memory ability levels, but also on the position of the item in the word list. This is not surprising, given the expected serial position effect observed.

CVLT-II Trial 5

As indicated in Figure 7, the serial position effect was also apparent on CVLT-II Trial 5 (F = 7.66, *p* < .01) utilizing a quadratic term (nonlinear quadratic regression). Notably, after repeated learning trials (CVLT-II Trials 2 to 5), the probability of recalling each item increased relative to CVLT-II Trial 1.

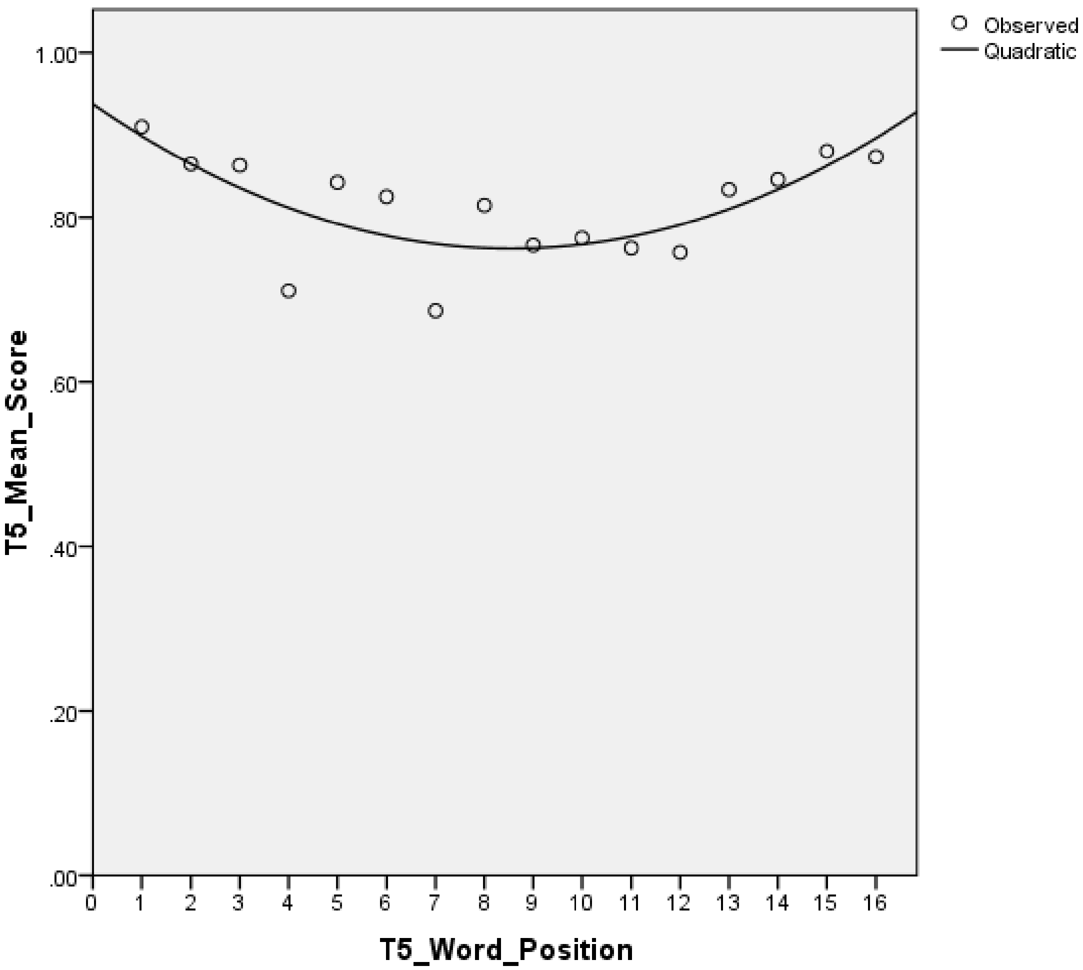


Figure 7 Regression model for CVLT-II Trial 5.

Data appear unidimensional for CVLT-II Trial 5, as demonstrated by the observation that there is only one eigenvalue from the real data > 1.0 in Figure 7. However, when comparing actual eigenvalues to those derived with modified parallel analysis data, the second eigenvalue of the observed data (0.94) is significantly larger than the second eigenvalue of the simulated data (0.78, *p* = .05). Therefore, modified parallel analysis results provided somewhat conflicting information regarding unidimensionality. As can be observed with a visual examination of the scree plot in Figure 8, data from CVLT-II Trial 5 appear unidimensional, and there is a clear, strong first factor. IRT analysis was therefore conducted.

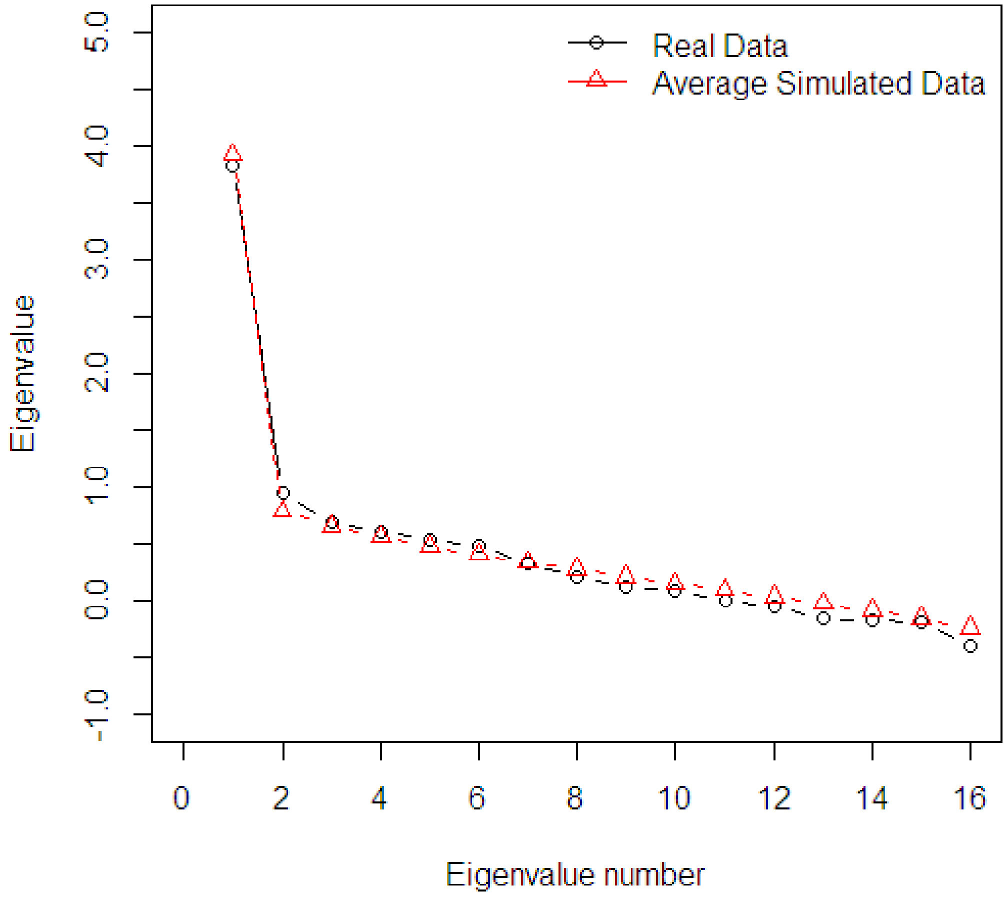


Figure 8 Scree plot for modified parallel analysis of CVLT-II Trial 5.

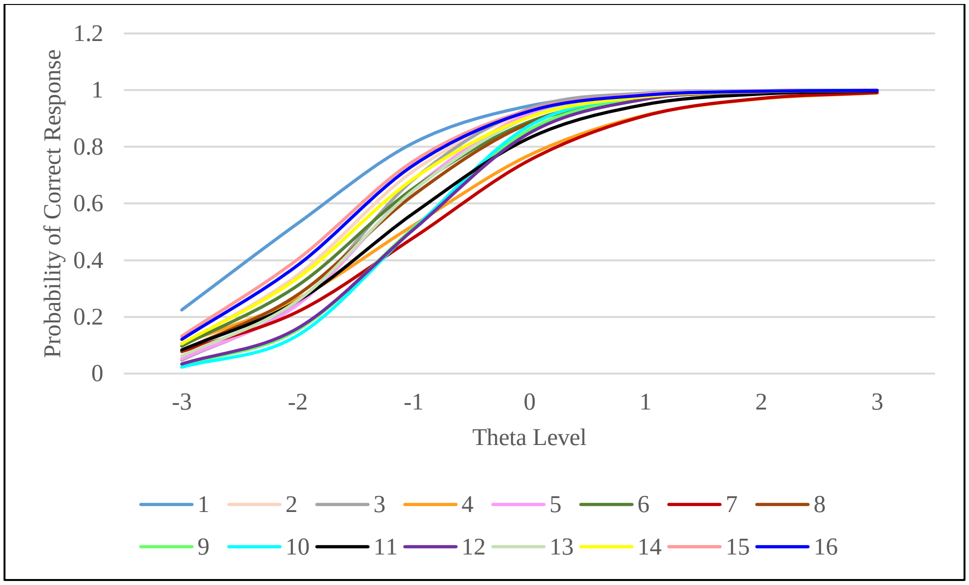
The proportion of examinees that answered each CVLT-II Trial 5 item correctly (P) is presented in Table 3. Findings from the 2 PL IRT analysis that provide estimates of item difficulty (β) and discrimination (α), as well as item fit statistics (Q_1 and *z*) are also presented in Table 3.

Table 3 IRT item parameters, Q_1, and *z* item fit statistics for CVLT-II Trial 5

Item	CVLT-II Trial 5					
	P	β	α	Q_1	*P*	Z
1	0.91	-2.09	0.80	16.33	<.01*	0.81
2	0.87	-1.60	0.91	9.01	.03*	0.55
3	0.86	-1.41	1.11	15.03	<.01*	0.33
4	0.71	-1.08	0.66	48.44	<.01*	0.85
5	0.84	-1.36	1.04	10.76	.01*	4.36*
6	0.83	-1.44	0.84	38.27	<.01*	0.71
7	0.69	-0.93	0.70	59.55	<.01*	0.96
8	0.82	-1.36	0.87	9.22	.03*	2.11*
9	0.77	-1.03	1.03	10.98	.01*	0.50
10	0.78	-1.02	1.12	11.96	<.01*	0.43
11	0.76	-1.20	0.78	44.70	<.01*	0.61
12	0.76	-1.02	0.99	11.10	.01*	0.50
13	0.83	-1.36	0.98	11.85	<.01*	0.39
14	0.85	-1.54	0.86	17.53	<.01*	0.59
15	0.88	-1.74	0.88	18.11	<.01*	0.61
16	0.87	-1.68	0.88	13.89	<.01*	0.56

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
**p* < .05.

In contrast to CVLT-II Trial 1, all CVLT-II Trial 5 items did not fit the IRT model according to Q_1 analyses. When evaluated with the standardized residual (*z*) fit statistic, 14 of 16 items appeared to fit the model. Given that two items (items 5 [onion] and 8 [zebra]) did not fit the model, subsequent analyses were conducted both with (results presented here), and without these two items (results presented in the next section). As indicated in Table 3, Item 10 (lamp) had the highest discrimination level (α = 1.12). Nevertheless, item difficulty and discrimination parameters are more evenly distributed in Trial 5, compared to Trial 1. For example, items in the middle of the list (average α = 0.92) had similar discrimination parameters with those of the first few (average α = 0.87) and last few (average α = 0.90) items, indicating that the items function rather similarly in terms of effectiveness of discriminating between examinees' verbal memory ability. Item characteristic curves, item information curves, and test information curve for all 16 items of CVLT-II Trial 5 are presented in Figures 9, 10, and 11 respectively. Monotonicity was evident across all 16 items of CVLT-II Trial 5. As the item characteristic and item information curves indicate, there is greater consistency in the way items function, compared to Trial 1. As the test information curve indicates, the maximum information that can be drawn from CVLT-II Trial 5 is approximately 9.20 at the theta level of -1.35. In summary, Trial 5 is more effective at discriminating between examinees with lower ability levels (θ = -1.35) compared to Trial 1 (θ = 0.10).



.

Figure 9 Item characteristic curves for CVLT-II Trial 5. Each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination.

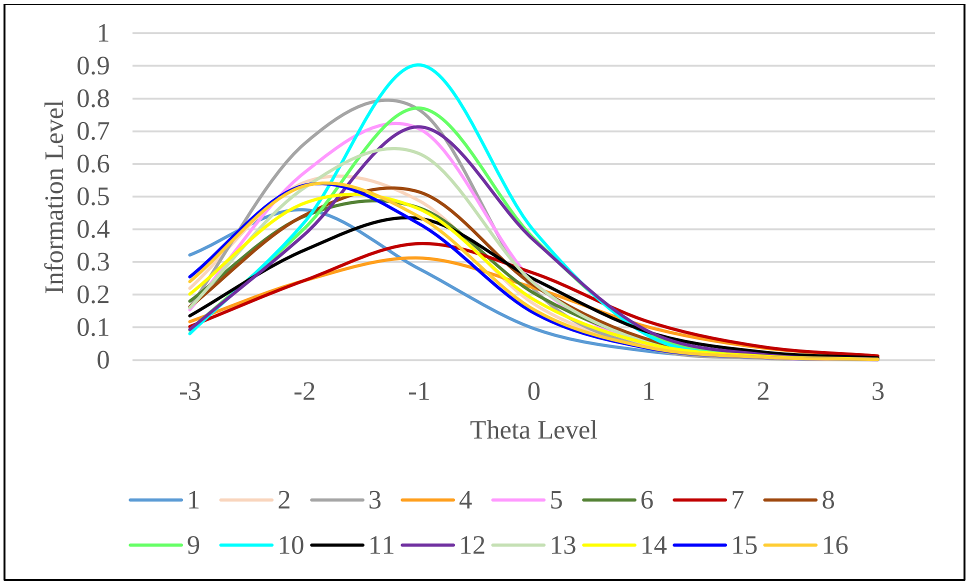


Figure 10 Item information curve for CVLT-II Trial 5. Each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent.

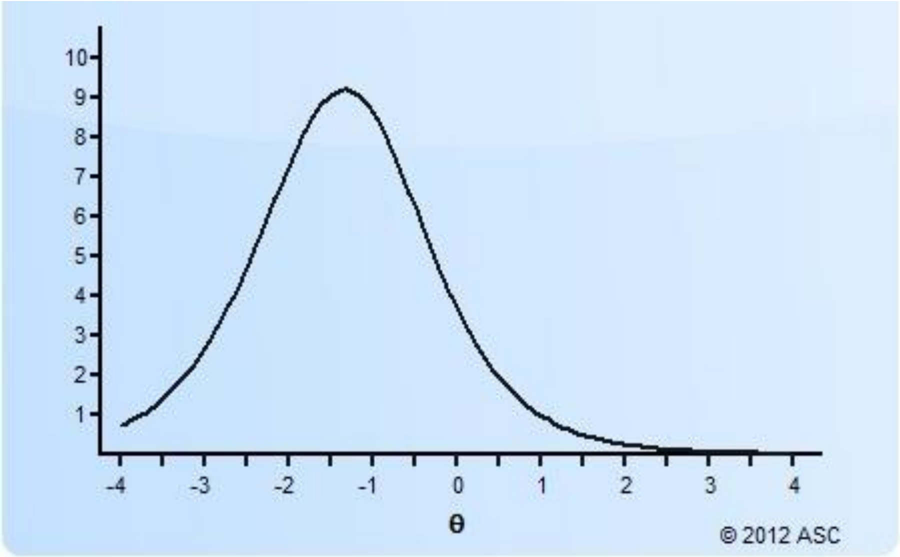


Figure 11 Test information curve for CVLT-II Trial 5. The curve represents the sum of item information at various ability levels, and indicates how well the test is estimating memory over the continuum of memory scores.

Correlations were small and non-significant between item position and item discrimination (*r* = .09, *p* = .74; r^2 = .06, *p* = .83). This indicates that local independence was assumed, and that discrimination levels for each of the 16 items in CVLT-II Trial 5 vary primarily based on verbal memory ability levels, independent of location of words on the list. Although the serial position effect was still evident (as shown in Figure 7 above), it was not so strong as Trial 1 to violate local independence.

CVLT-II Trial 5, 14 items

Given that two items (items 5 [onion] and 8 [zebra]) from CVLT-II Trial 5 did not fit the IRT model, even when evaluated using standardized residual (*z*) item fit statistics, analyses were conducted with items 5 and 8 excluded.

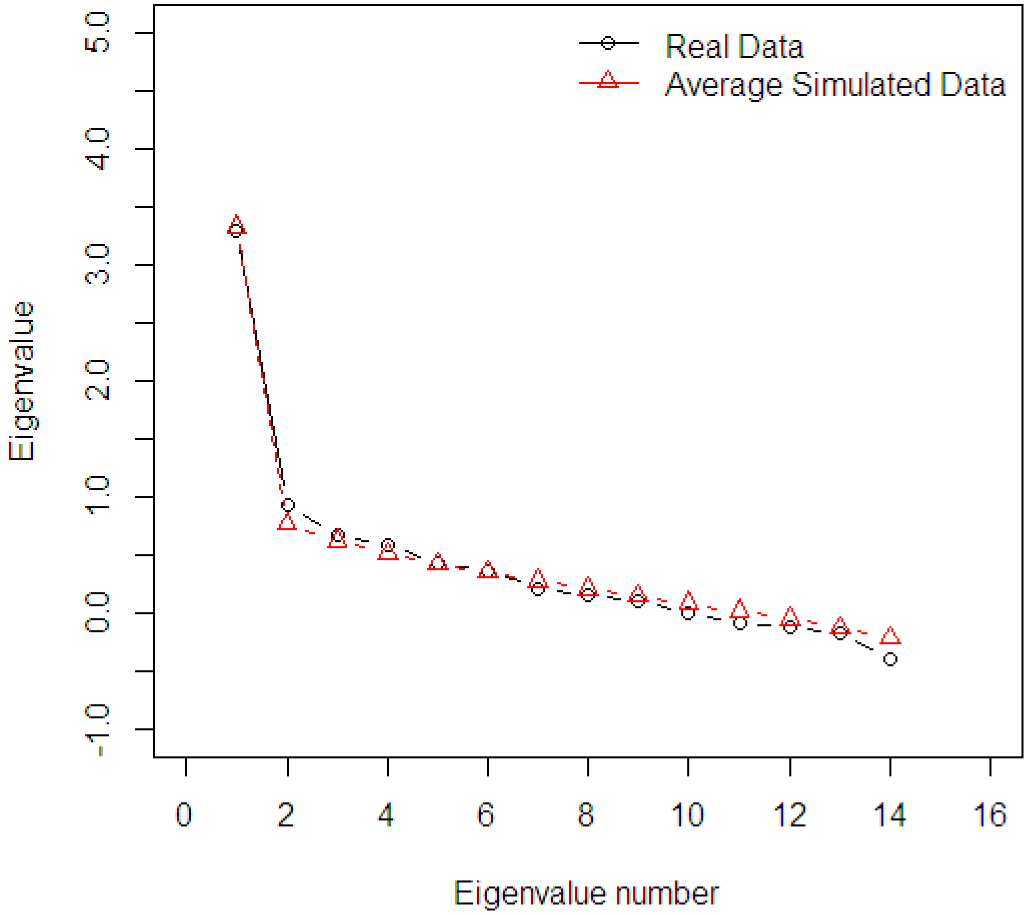


Figure 12 Scree plot for modified parallel analyses of CVLT-II Trial 5 (14-items).

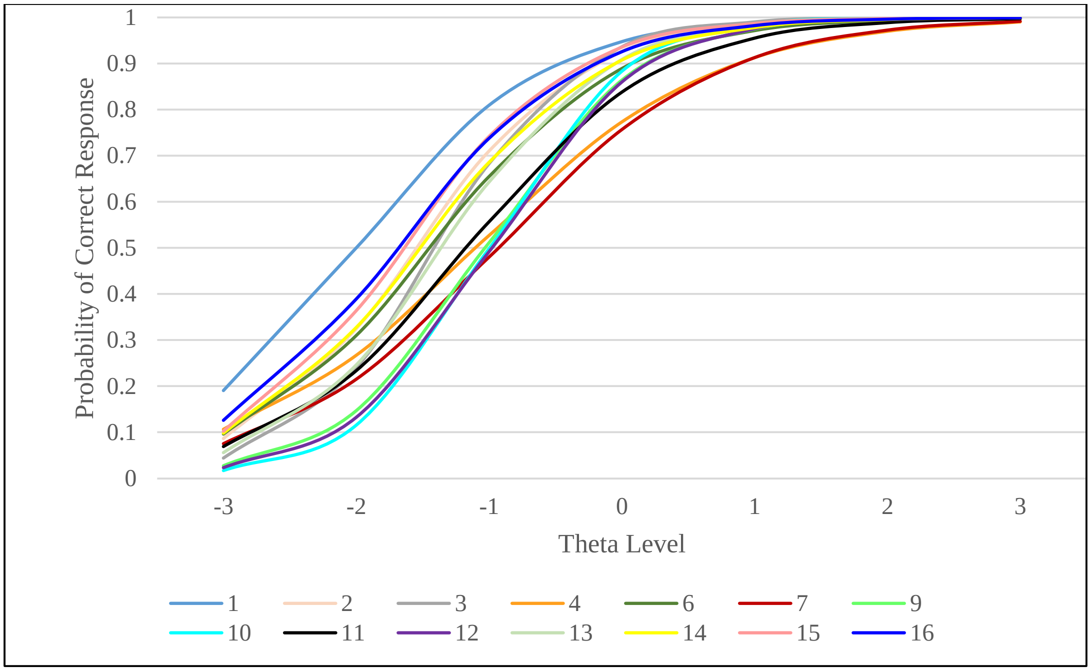
As can be observed in Figure 12, modified parallel analysis again did not indicate unidimensionality, but visual inspection of the scree plot indicated that the model appears to be unidimensional. Removal of Items 5 and 8 did not appear to significantly affect the degree to which data was unidimensional. The proportion of examinees that answered each item correctly (P), item difficulty (β) and discrimination (α) parameters, as well as item fit statistics (Q_1 and *z*) are presented in Table 4.

Table 4 IRT item parameters, Q_1, and *z* item fit statistics for CVLT-II Trial 5 (14-items)

Item	CVLT-II Trial 5					
	P	β	α	Q_1	*p*	Z
1	0.91	-2.00	0.85	14.52	<.01*	0.76
2	0.87	-1.55	0.96	12.29	<.01*	0.56
3	0.86	-1.40	1.13	16.74	<.01*	0.39
4	0.71	-1.10	0.66	106.07	<.01*	0.95
6	0.83	-1.44	0.85	19.03	<.01*	0.78
7	0.69	-0.93	0.71	162.19	<.01*	1.00
9	0.77	-1.02	1.06	22.93	<.01*	0.63
10	0.78	-1.00	1.19	13.95	<.01*	0.39
11	0.76	-1.16	0.83	18.77	<.01*	0.69
12	0.76	-0.98	1.08	11.82	<.01*	0.53
13	0.83	-1.34	1.00	15.96	<.01*	0.51
14	0.85	-1.52	0.88	22.02	<.01*	0.62
15	0.88	-1.65	0.95	22.88	<.01*	0.51
16	0.87	-1.70	0.87	14.68	<.01*	0.68

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
*$p < .05$.

Not surprisingly, all 14 CVLT-II Trial 5 items included in these analyses did not fit an IRT model according to Q_1 analyses. However, all 14 items appeared to fit the model, according to the standardized residual (*z*) statistics. Removal of the two items did not significantly change item difficulty or discrimination parameters of the remaining 14 items in CVLT-II Trial 5. Item characteristic curves, item information curves, and test information curve for CVLT-II Trial 5 (14-item) are presented in Figures 13, 14, and 15, respectively. Monotonicity was evident across all 14 items of CVLT-II Trial 5 (14-item), and the maximum information that can be drawn from CVLT-II Trial 5 (14-item) is approximately 8.49 at the theta level of -1.30, which is similar to that of Trial 5 (16 items), whereby maximum information was drawn at a theta level of -1.35.



.

Figure 13 Item characteristic curves for CVLT-II Trial 5 (14-item; excluding items 5 and 8). Each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination.

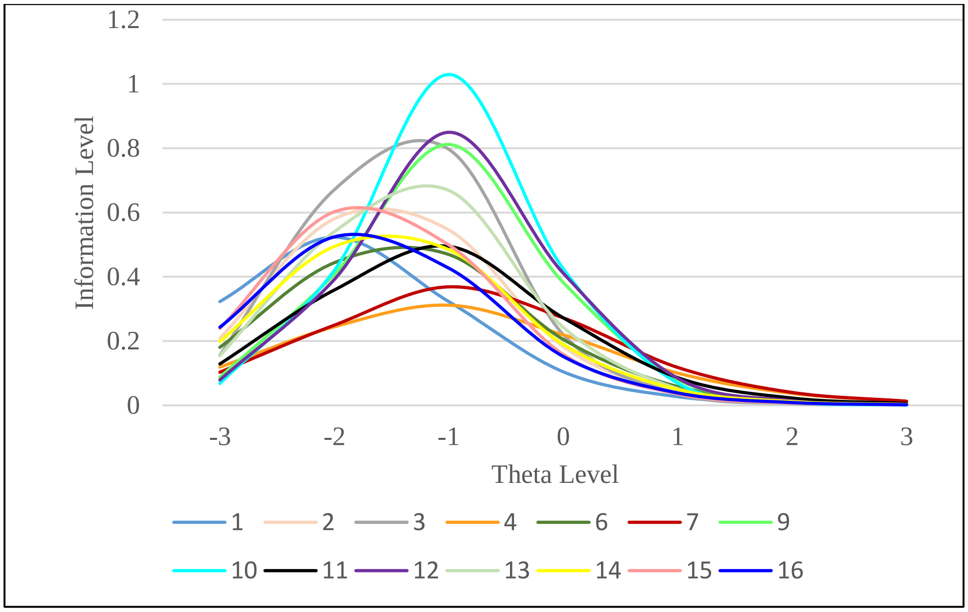


Figure 14 Item information curves for CVLT-II Trial 5 (14-item; excluding items 5 and 8). Each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent.

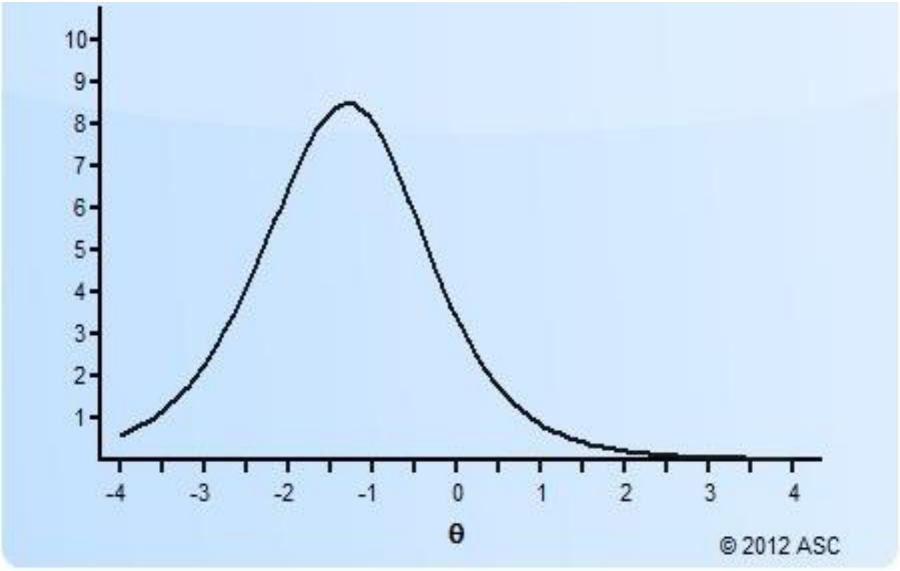


Figure 15 Test information curve for CVLT-II Trial 5 (14-item). The curve represents the sum of item information at various ability levels, and indicates how well the test is estimating memory over the continuum of memory scores.

Similar to those of the 16-item version of CVLT-II Trial 5, the parameter estimates for the 14-item version of CVLT-II Trial 5 are likely unbiased, given that local independence was assumed, as observed by small and non-significant correlations between item position and item discrimination (r = .15, p = .61; r^2 = .12, p = .67). Overall, removing the two items did not appear to significantly change the degree to which the model met the assumptions of unidimensionality, monotonicity, or local independence.

CVLT-II Short Delay Free Recall

To investigate the serial position effect, mean scores for all 16 items in CVLT-II Trial SDFR were regressed onto the item sequence number (1 to 16). As indicated in Figure 16, the serial position effect was not apparent on CVLT-II Trial SDFR (F = 17.54, p < .01), whereby a linear term seemed to best fit the data. As the graph depicts, examinees tended to recall more of the items presented earlier in the word list, and less of the items presented later in the word list. Thus, there is evidence of a primacy effect, but not a recency effect.

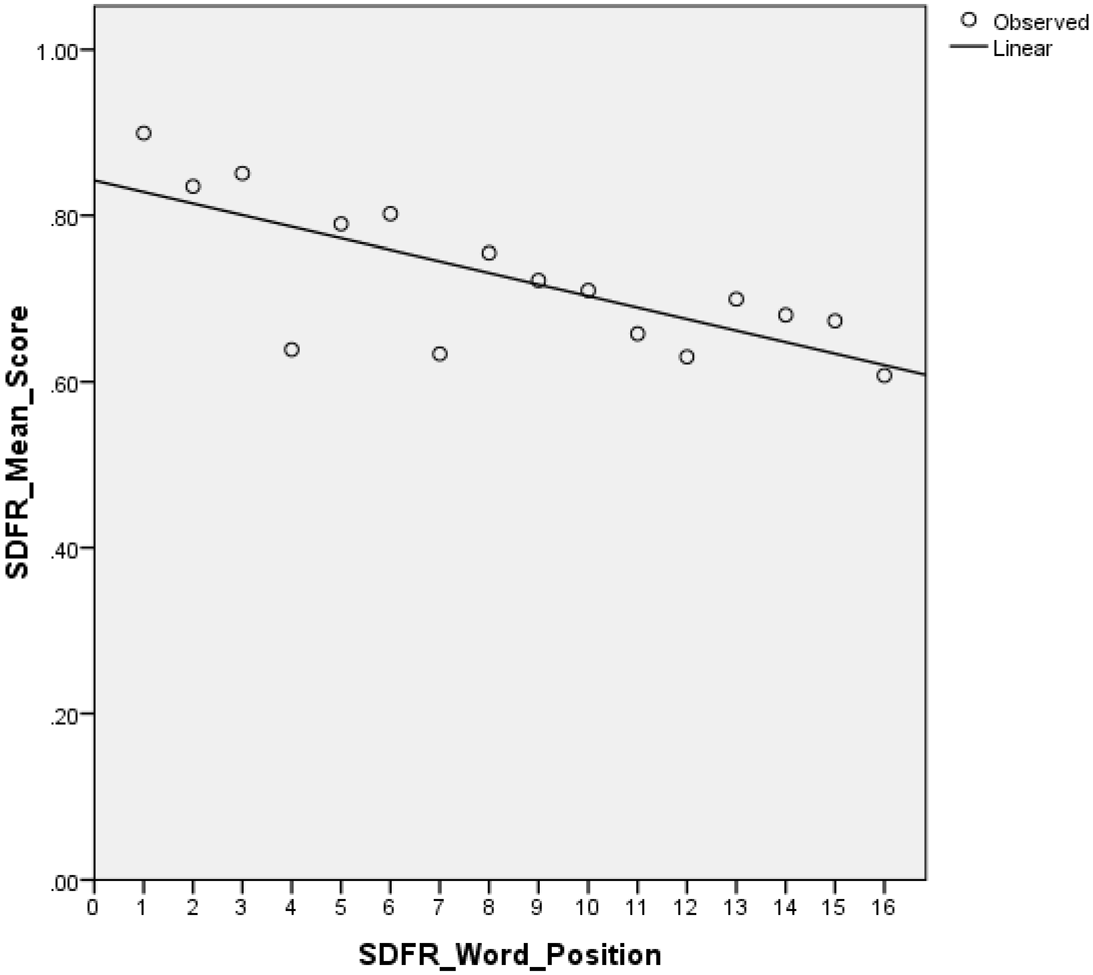


Figure 16 Regression model for CVLT-II Trial SDFR.

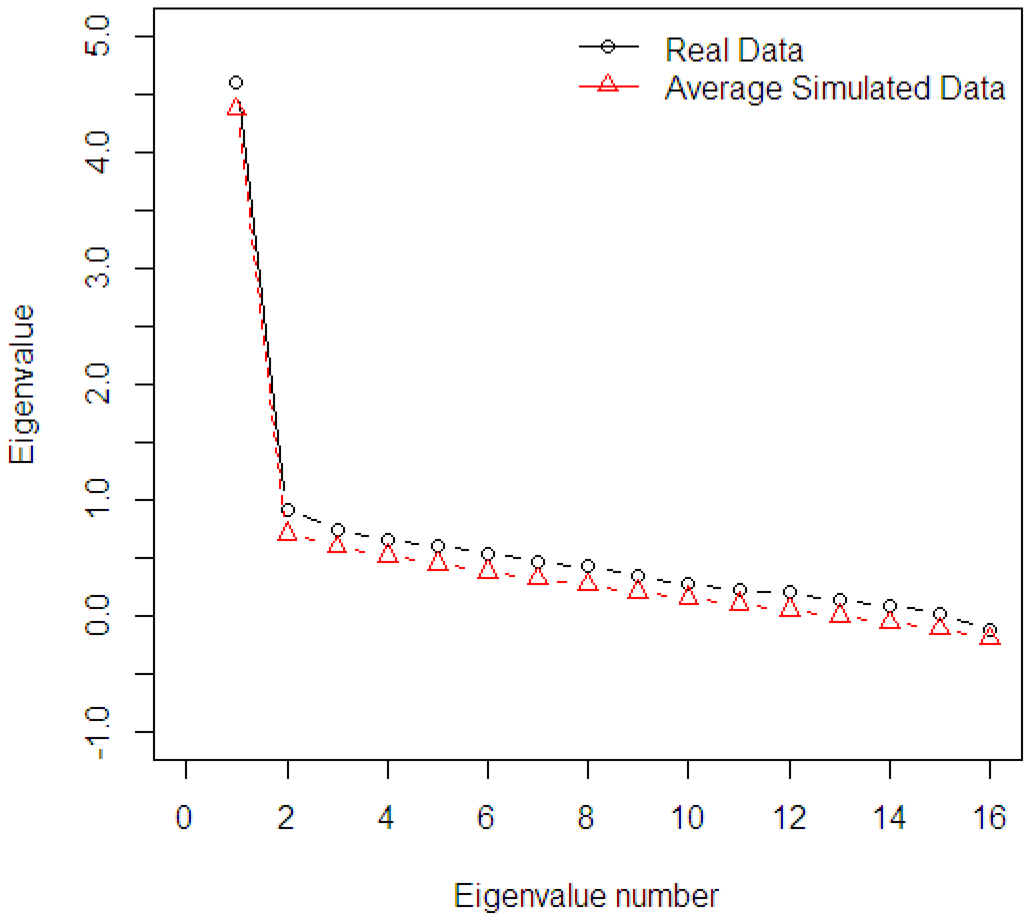


Figure 17 Scree plot for modified parallel analysis of CVLT-II Trial SDFR.

As can be observed with a visual examination of the scree plot in Figure 17, data appear unidimensional, and IRT analysis was therefore conducted. The proportion of examinees that answered each item correctly (P), estimates of item difficulty (β) and discrimination (α) parameters, as well as item fit statistics (Q_1 and *z*) are presented in Table 5.

Table 5 IRT item parameters, Q_1, and *z* item fit statistics for CVLT-II Trial SDFR

Item	CVLT-II Trial SDFR					
	P	β	α	Q_1	*p*	*z*
1	0.90	-1.80	0.93	7.90	.05*	0.35
2	0.84	-1.55	0.77	11.29	.01*	0.64
3	0.85	-1.38	1.04	7.86	.05*	0.23
4	0.64	-0.70	0.66	22.81	<.01*	0.54
5	0.79	-1.34	0.73	10.21	.02*	0.66
6	0.80	-1.49	0.68	15.72	<.01*	0.81
7	0.63	-0.74	0.57	16.07	<.01*	0.81
8	0.76	-1.06	0.86	20.50	<.01*	0.56
9	0.72	-0.92	0.85	8.85	.03*	0.40
10	0.71	-0.74	1.16	8.43	.04*	0.42
11	0.66	-0.72	0.76	11.71	<.01*	0.66
12	0.63	-0.53	0.98	14.19	<.01*	0.56
13	0.70	-0.78	0.95	17.07	<.01*	0.50
14	0.68	-0.77	0.83	8.92	.03*	0.46
15	0.67	-0.72	0.87	6.33	.10	0.39
16	0.61	-0.51	0.78	13.08	<.01*	0.54

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
$^*p < .05$.

Fifteen of 16 items in CVLT-II Trial SDFR did not fit the IRT model according to Q_1 analyses, but all 16 items appeared to fit the model, according to the standardized residual (*z*) statistics. Similar to the estimates derived from Trial 5, item difficulty and discrimination estimates were more uniformly distributed among the 16 items in SDFR, compared to Trial 1. As in CVLT-II Trial 5, Item 10 continued to have the highest discrimination estimate (α = 1.16). Item characteristic curves, item information curves, and test information curve for CVLT-II Trial SDFR are presented in Figures 18, 19, and 20, respectively. Monotonicity was evident across all 16 items of CVLT-II Trial SDFR, and the maximum information that can be drawn from CVLT-II Trial SDFR is approximately 7.74 at the theta level of -0.95.

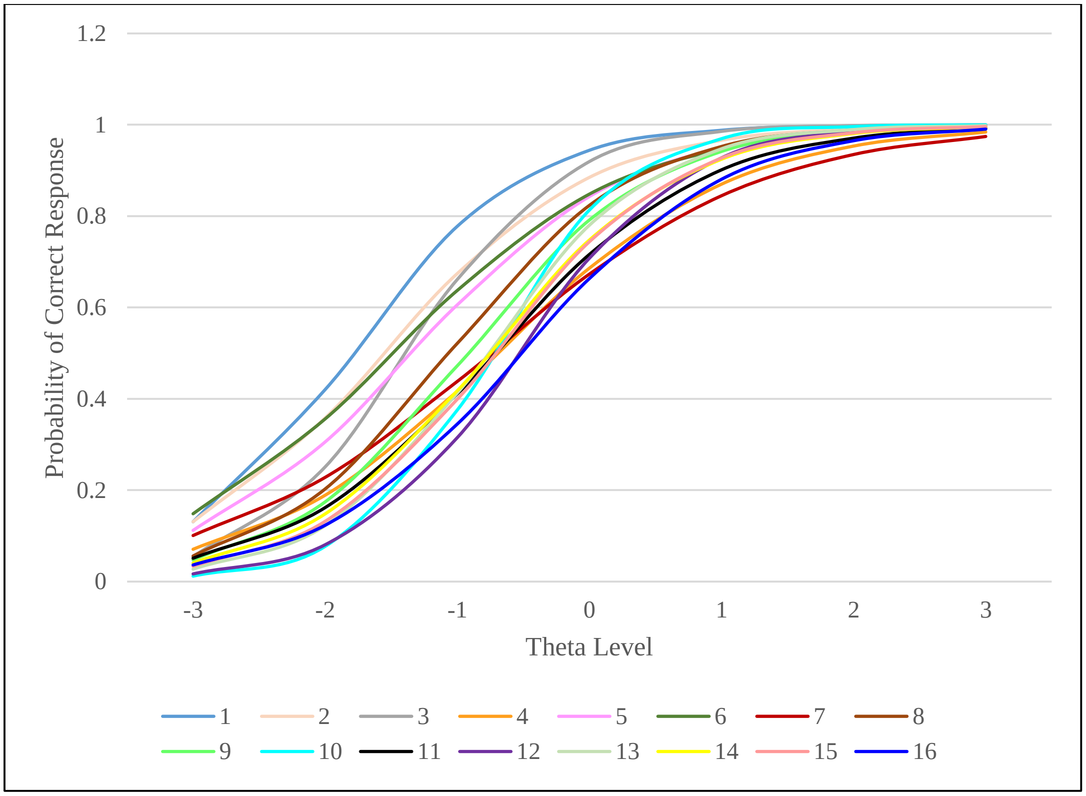


Figure 18 Item characteristic curves for CVLT-II Trial SDFR. Each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination.

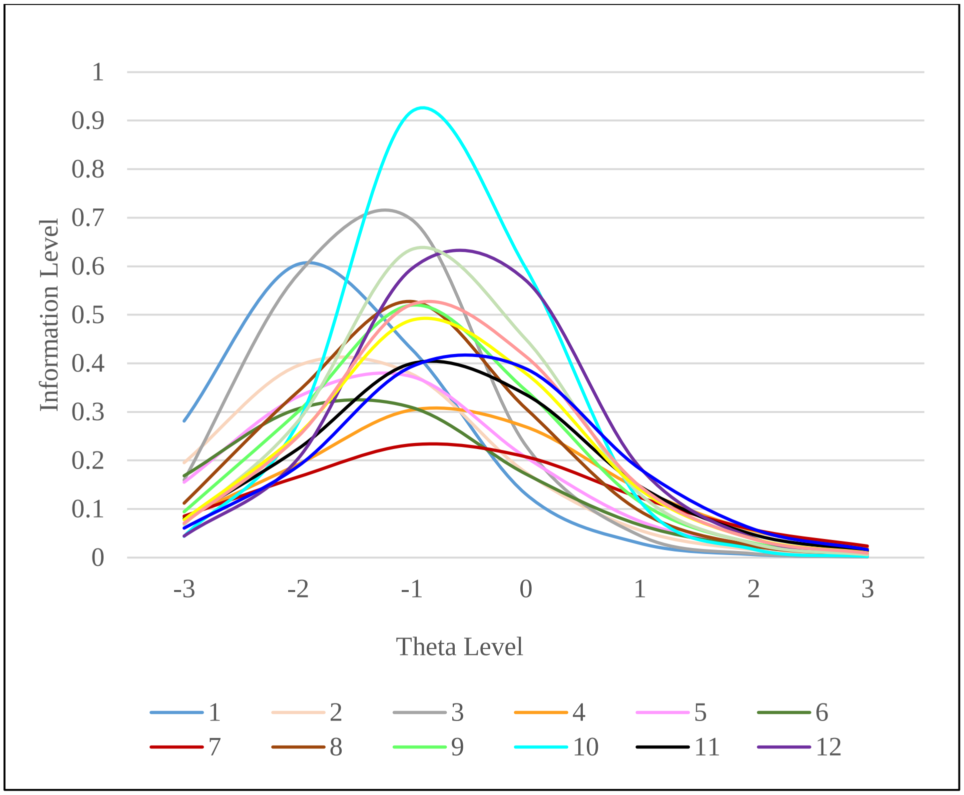


Figure 19 Item Information curves for CVLT-II Trial SDFR. Each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent.

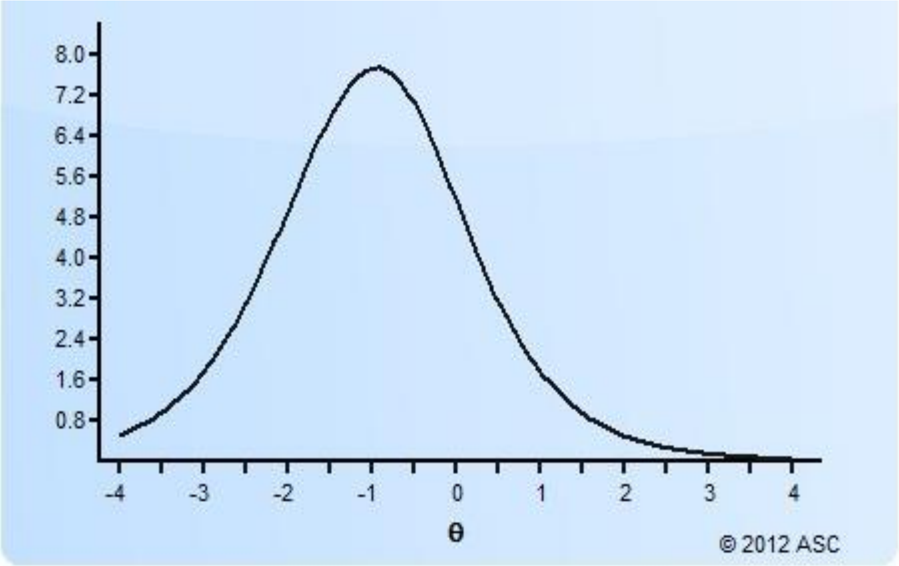


Figure 20 Information curve for CVLT-II Trial SDFR. The curve represents the sum of item information at various ability levels, and indicates how well the test is estimating memory over the continuum of memory scores.

Correlations among item position, item discrimination, and squared item discrimination were small and non-significant (*r* = .15, *p* = .58; r^2 = .13, *p* = .64), thus indicating that parameter estimates for CVLT-II SDFR are likely unbiased.

CVLT-II Long Delay Free Recall

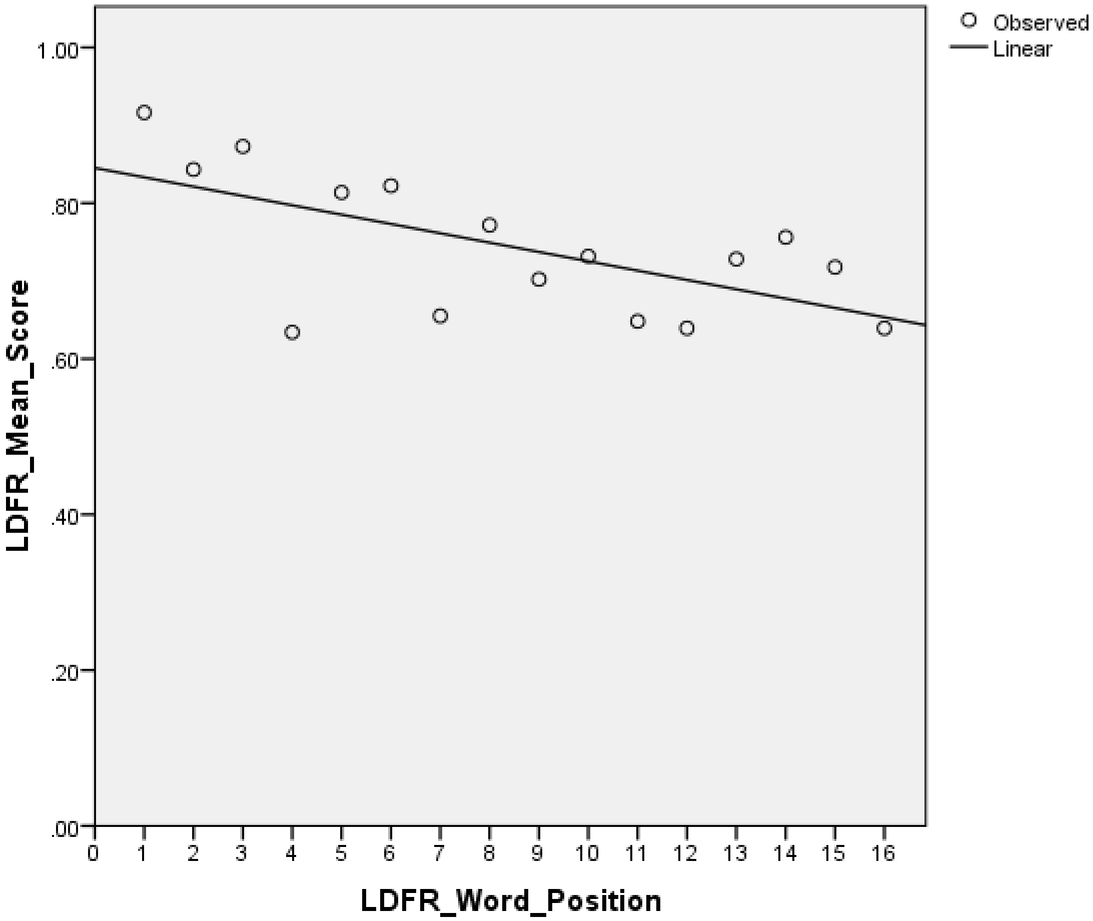


Figure 21 Regression model for CVLT-II Trial LDFR.

As indicated in Figure 21, the serial position effect was less apparent on CVLT-II Trial LDFR (F = 9.36, *p* < .01), whereby a linear term seemed to best fit the data. Similar to the pattern observed in CVLT-II Trial SDFR, this suggests that examinees tended to remember less of the words presented later in the list, and more of the words presented earlier in the list.

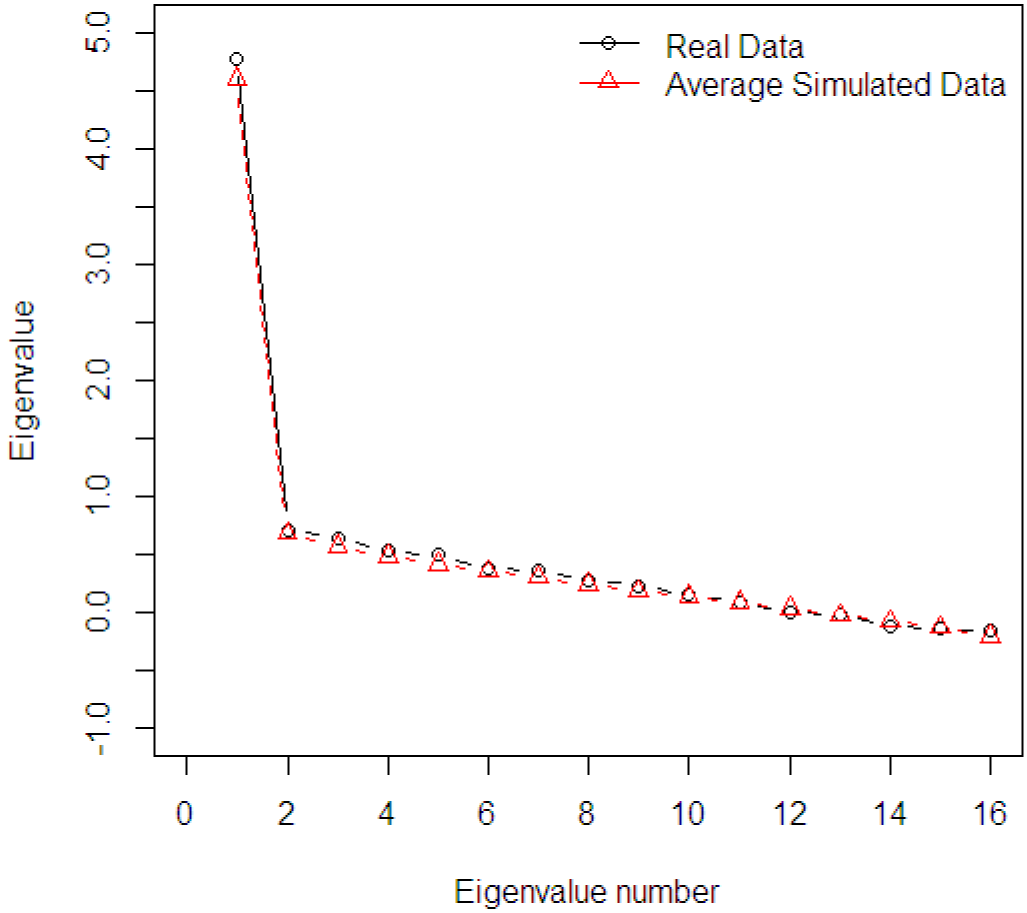


Figure 22 Scree plot for modified parallel analysis of CVLT-II Trial LDFR.

As can be observed with a visual examination of the scree plot in Figure 22, data appear unidimensional. The proportion of examinees that answered each item correctly

(P), estimates of item difficulty (β) and discrimination (α), as well as item fit statistics (Q_1 and *z*) are presented in Table 6.

Table 6 IRT item parameters, Q_1, and *z* item fit statistics for CVLT-II Trial LDFR

Item	CVLT-II Trial LDFR					
	P	β	α	Q_1	*p*	*z*
1	0.92	-1.85	1.00	16.13	<.01*	0.53
2	0.84	-1.43	0.93	17.36	<.01*	0.79
3	0.87	-1.56	0.98	18.65	<.01*	0.39
4	0.63	-0.73	0.60	26.05	<.01*	0.78
5	0.81	-1.42	0.78	23.70	<.01*	0.90
6	0.82	-1.44	0.81	25.60	<.01*	0.64
7	0.66	-0.67	0.85	36.13	<.01*	0.60
8	0.77	-1.06	0.96	12.49	<.01*	0.39
9	0.70	-0.83	0.88	12.15	<.01*	0.53
10	0.73	-0.89	0.98	20.02	<.01*	0.36
11	0.65	-0.64	0.86	19.43	<.01*	0.64
12	0.64	-0.54	1.10	38.24	<.01*	0.71
13	0.73	-0.88	0.97	12.07	<.01*	0.33
14	0.76	-1.06	0.87	28.54	<.01*	0.65
15	0.72	-0.88	0.90	23.63	<.01*	0.32
16	0.64	-0.63	0.80	22.48	<.01*	0.78

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
*p < .05.

All 16 items in CVLT-II Trial LDFR did not fit the IRT model according to Q_1 analyses but when evaluated with the standardized residual (*z*) statistic, all items appeared to fit the model. Similar to the estimates derived from Trial 5 and SDFR, item difficulty and discrimination estimates were more uniformly distributed among the 16 items in SDFR, compared to Trial 1. Item 12 (cow) had the highest discrimination level (α = 1.10) and difficulty level (β = -0.54). Item characteristic curves, item information curves, and test information curve for CVLT-II Trial LDFR are presented in Figures 23, 24, and 25, respectively. Monotonicity was evident across all 16 items of CVLT-II Trial LDFR, and the maximum information that can be drawn from CVLT-II Trial LDFR is approximately 8.52 at theta = -1.00. Correlations among item position, item discrimination, and squared item discrimination were small and non-significant (*r* = .09, *p* = .73; r^2 = .07, *p* = .79), indicating that item parameter estimates are unbiased.

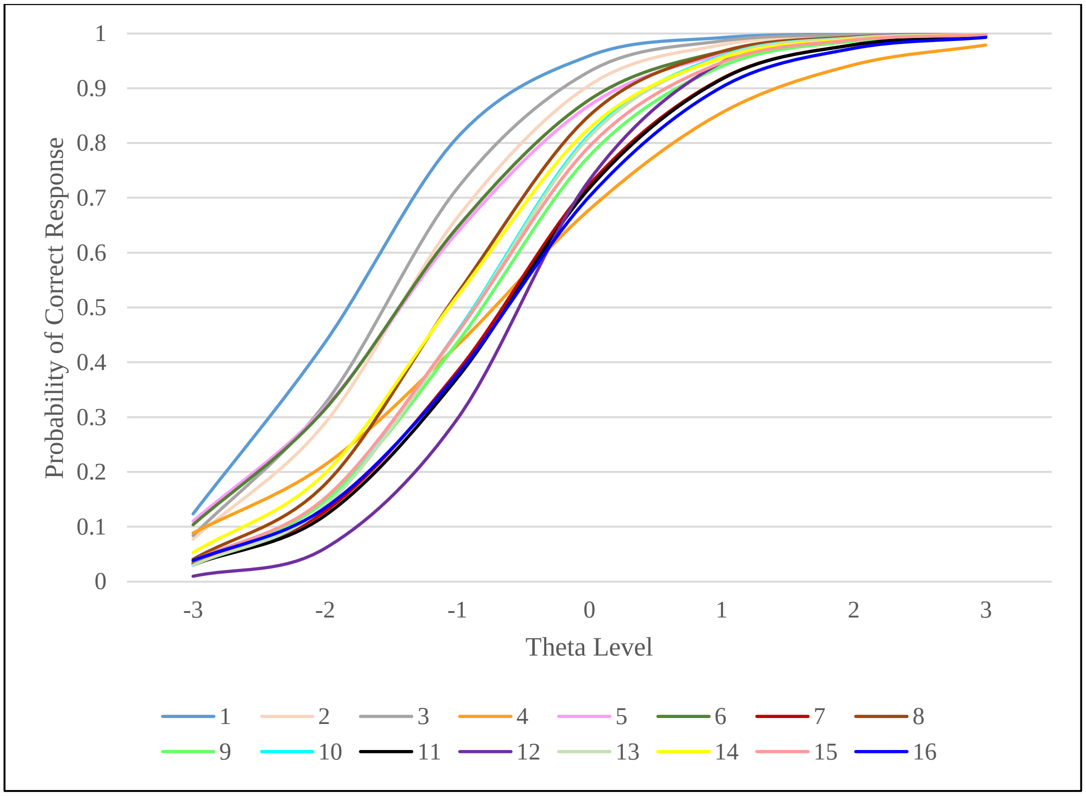


Figure 23 Item characteristic curves for CVLT-II Trial LDFR. Each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination.

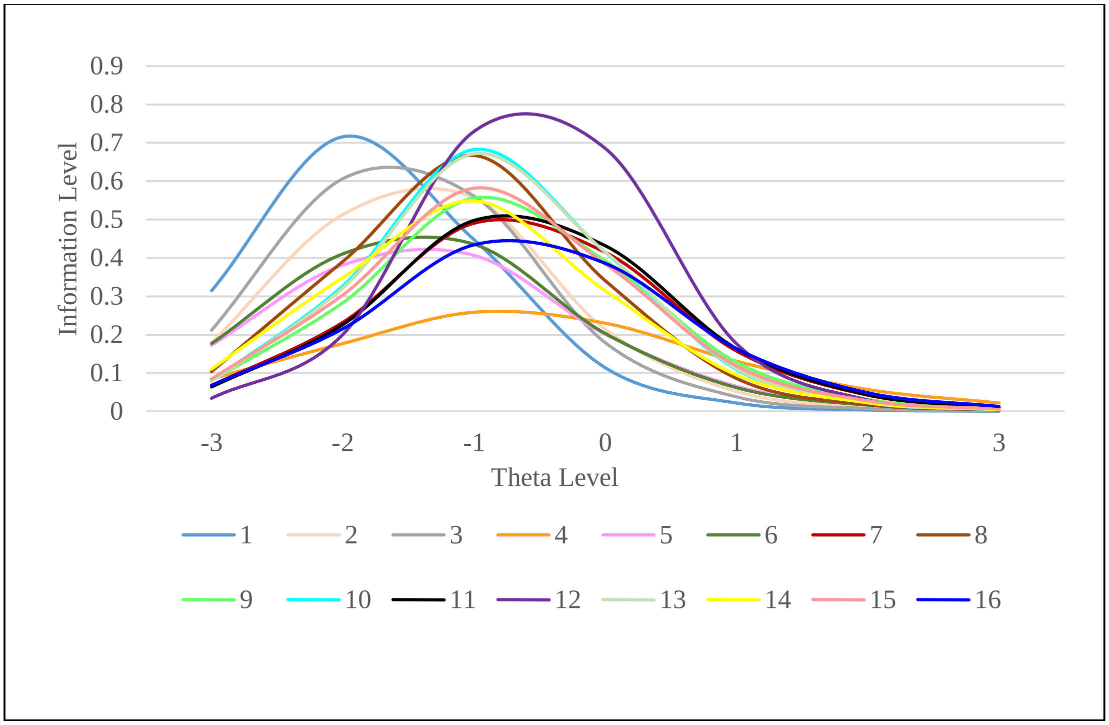


Figure 24 Item information curves for CVLT-II Trial LDFR. Each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent.

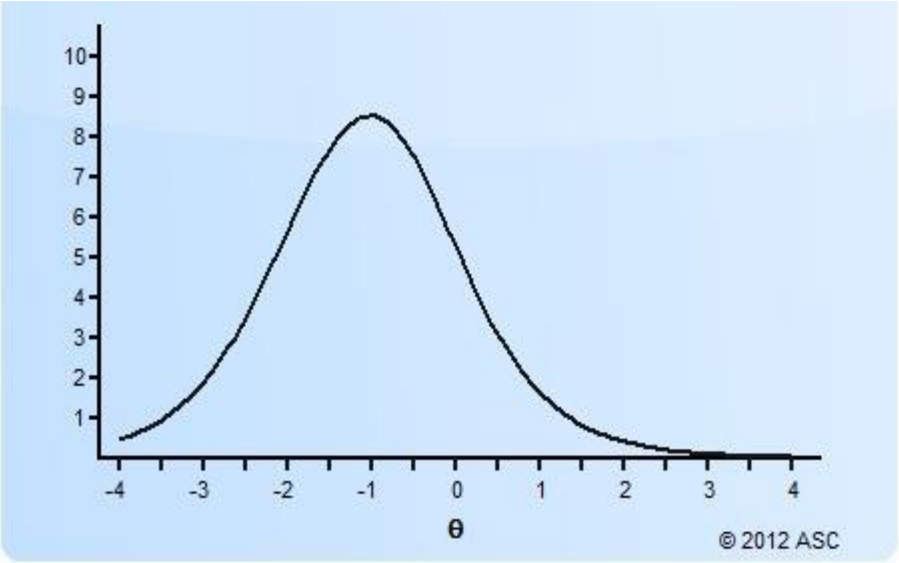


Figure 25 Test information curve for CVLT-II Trial LDFR. The curve represents the sum of item information at various ability levels, and indicates how well the test is estimating memory over the continuum of memory scores.

Summary of IRT Results for CVLT-II Trials

Overall, Trial 1 (mean difficulty level = 0.67) had the highest difficulty parameters of the four trials investigated (mean difficulty level of Trial 5 = -1.37, SDFR = -0.99, LDFR = -1.03). With repeated exposure to the same word list, recall of items improved (i.e., the test became easier), and the measure appears most effective at discriminating between examinees at the lower end of the memory ability spectrum. However, Trial 1 also had the lowest mean discrimination level (0.38) of the four trials, and mean discrimination levels were similar across Trials 5 (0.90), SDFR (0.84), and LDFR (0.89). In other words, Trial 1 was not very effective at discriminating between examinees, but the degree to which the test was able to distinguish between examinees with different ability levels increased in Trials 5, SDFR, and LDFR. The item with the highest difficulty level within each trial was located in the middle of the list. Similarly, with the exception of Trial 1, the item with the highest discrimination value within a list was also derived from the middle portion of the list. These findings suggest that items differ in their effectiveness in discriminating between examinees across the memory ability continuum.

LM I Story B

The serial position effect was not apparent in LM I Story B, as indicated in Figure 26, whereby a cubic term (F = 3.62, *p* = .03) seemed to best fit the data. This suggests that examinees most frequently recalled information from the beginning of the story, and other details from the middle of the story.

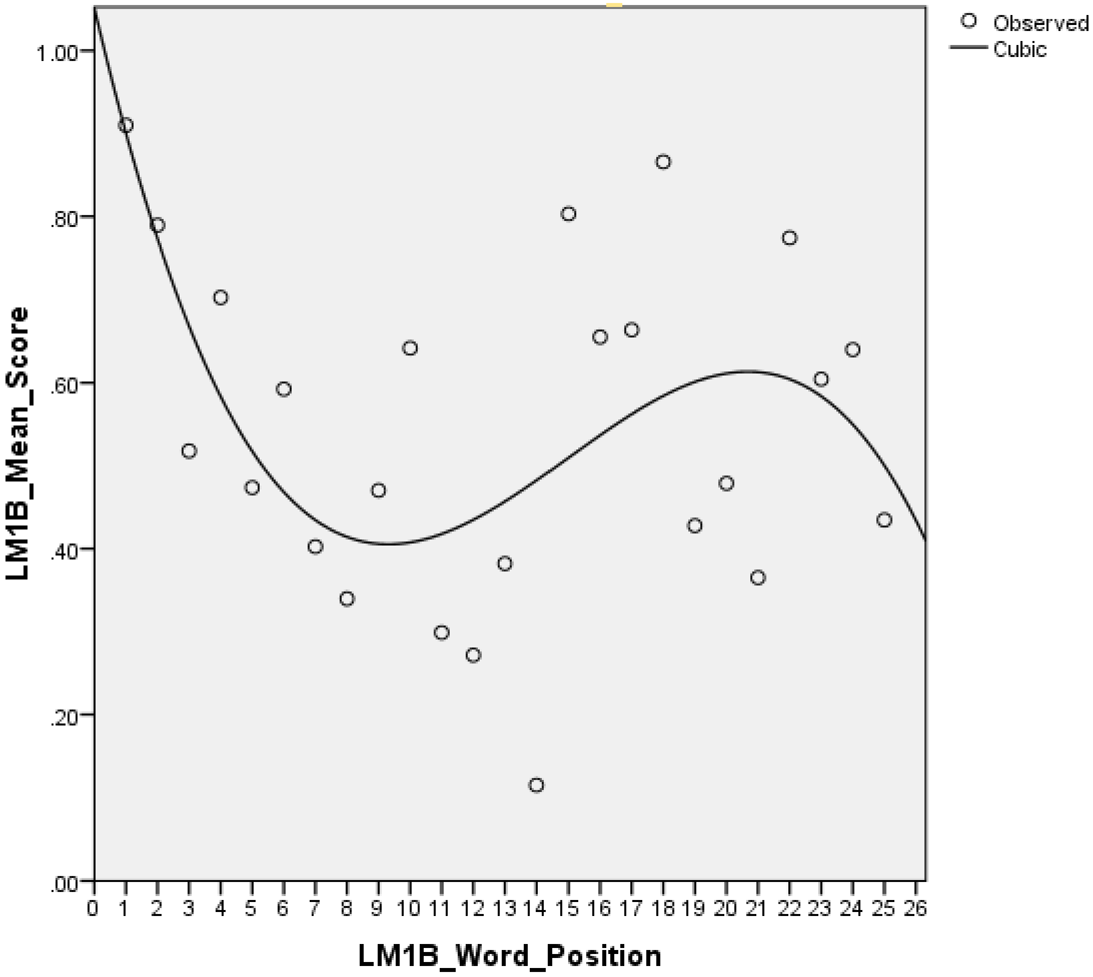


Figure 26 Regression model for LM I Story B.

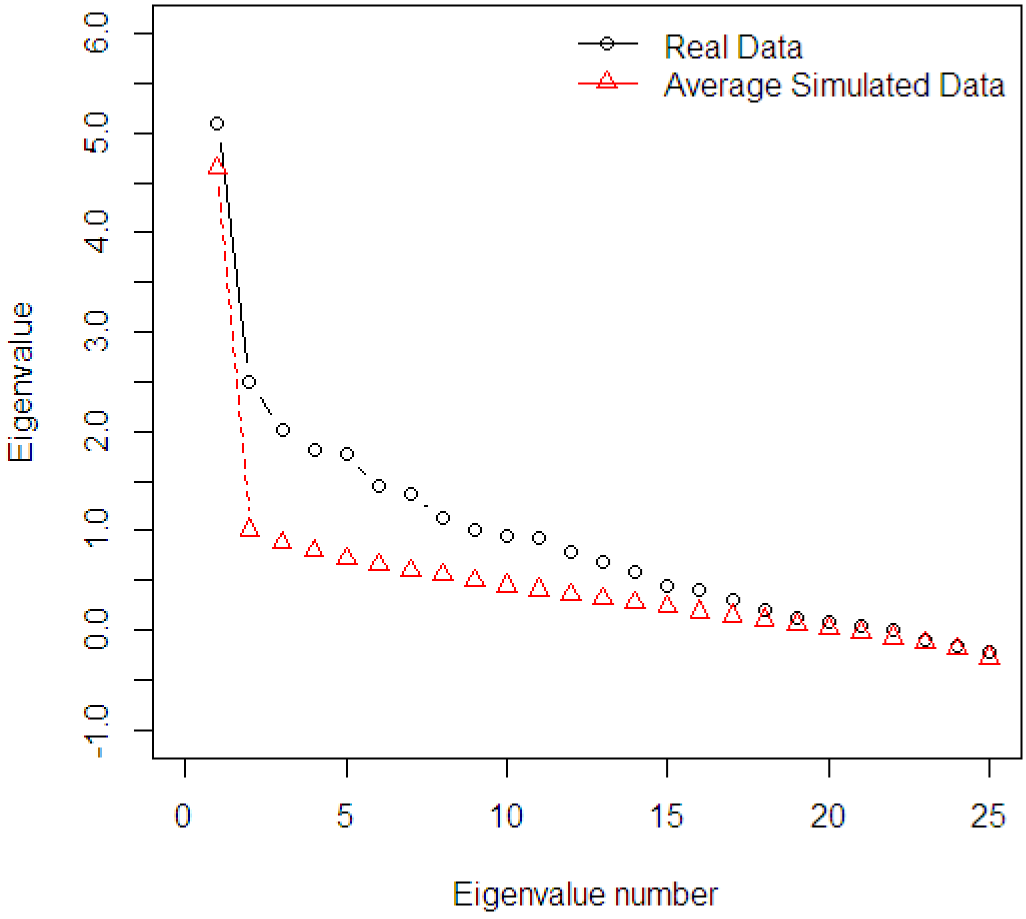


Figure 27 Scree plot for modified parallel analysis of LM I Story B.

As can be observed in Figure 27, there were two eigenvalues > 1 from the observed data, suggesting a potential violation of the unidimensionality assumption. However, a visual examination of the scree plot indicated that the violation of this assumption is not so severe as to negate the use of IRT. The proportion of examinees that answered each item correctly (P), estimates of item difficulty (β) and discrimination (α), as well as item fit statistics (Q_1 and *z*) are presented in Table 7.

Table 7 IRT item parameters, Q_1, and *z* item fit statistics for LM I Story B

Item	LM I Story B					
	P	β	α	Q_1	*p*	*z*
1	0.91	-2.60	0.58	6.95	.07	0.44
2	0.79	-1.54	0.57	5.04	.17	0.18
3	0.52	-0.10	0.61	4.57	.21	0.31
4	0.70	-1.17	0.48	4.29	.23	0.44
5	0.47	0.15	0.40	6.03	.11	0.32
6	0.59	-0.55	0.44	4.02	.26	0.41
7	0.40	0.49	0.51	9.68	.02*	0.35
8	0.34	0.77	0.56	9.14	.03*	0.28
9	0.47	0.14	0.48	8.41	.04*	0.29
10	0.64	-1.12	0.32	9.98	.02*	0.85
11	0.30	1.50	0.34	13.68	<.01*	0.93
12	0.27	1.65	0.36	8.79	.03*	0.80
13	0.38	0.72	0.41	13.10	<.01*	0.58
14	0.12	2.31	0.56	6.21	.10	0.44
15	0.80	-1.64	0.57	4.92	.18	0.67
16	0.66	-0.77	0.57	6.62	.09	0.44
17	0.66	-0.75	0.64	8.13	.04*	0.38
18	0.87	-1.71	0.79	4.63	.20	0.27
19	0.43	0.35	0.53	5.96	.11	0.49
20	0.48	0.08	0.60	5.35	.15	0.41
21	0.37	0.58	0.64	2.40	.49	0.28
22	0.77	-1.14	0.81	4.48	.21	0.47
23	0.60	-0.53	0.55	3.50	.32	0.28
24	0.64	-0.56	0.79	0.62	.89	0.31
25	0.44	0.31	0.54	4.23	.24	0.31

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
**p* < .05.

Eight of the 25 items in LM I Story B did not fit the IRT model according to Q_1 analyses, but all 25 items appeared to fit the model, according to standardized residual (*z*) statistics. The first two items (first and last name of the main character in the story), as well as several items in the middle of the story (e.g., items 10, 15, 18, and 22) appeared to have very low difficulty levels, arguably related to distinctive details of the story (e.g., Item 10 = police, item 15 = robbed, 18 = children, item 22 = police). Item discrimination estimate was highest for item 22 (police; α =0.81), but generally appeared quite dispersed across all 25 items (α range from 0.32 to 0.79). Item characteristic curves, item information curves, and test information curve for LM I Story B are presented in Figures 28, 29, and 30 respectively. Monotonicity was evident across all 25 items of LM I Story B, and the maximum information that can be drawn from LM I Story B is approximately 4.59 at the theta level of -0.60. Correlations among item position, item discrimination, and squared item discrimination were moderate and significant (r = .46, p = .02; r^2 = .48, p = .01), indicating that parameter estimates for LM I Story B are likely slightly biased. In other words, the degree to which an item discriminates between different ability levels is not solely dependent on verbal memory ability levels, but also on the position of the item and/or distinctive content of the item in the story.

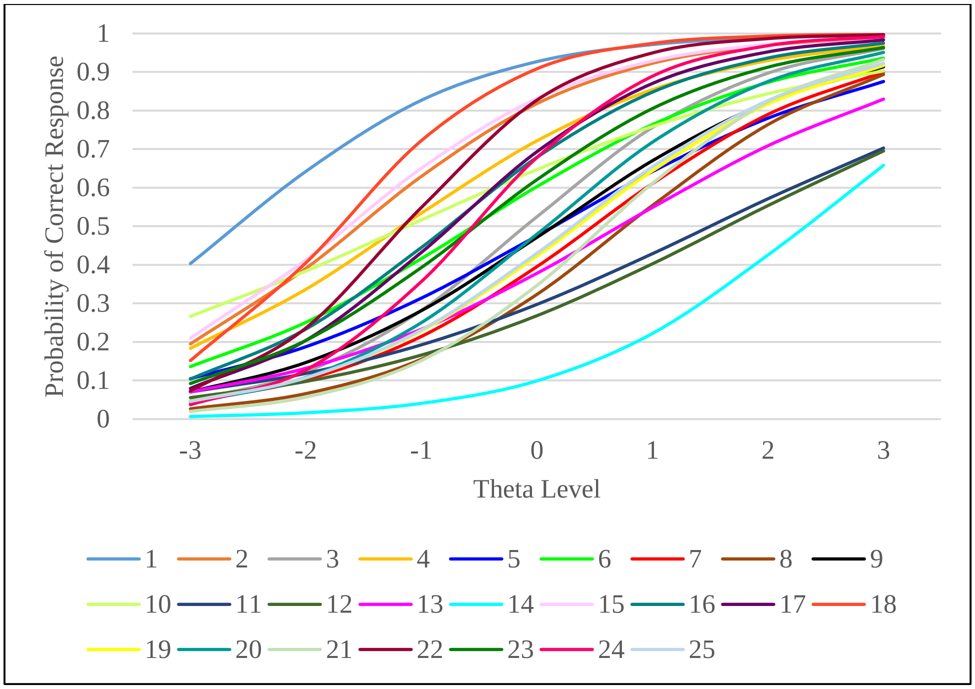


Figure 28 Item characteristic curves for LM I Story B. Each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination.

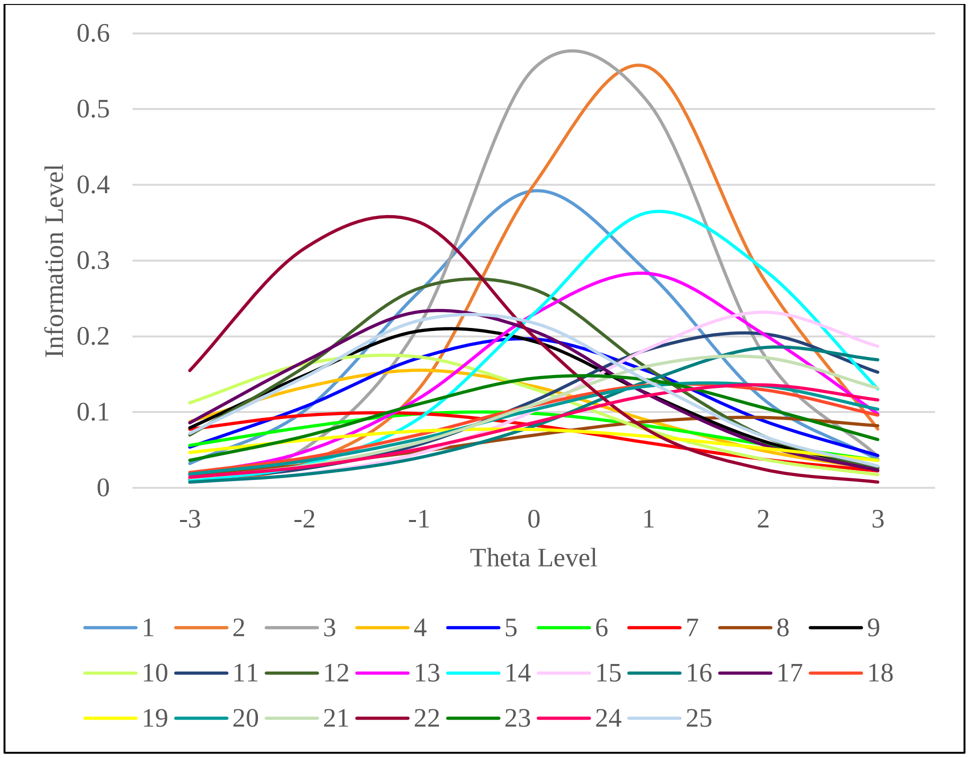


Figure 29 Item information curves for LM I Story B. Each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent.

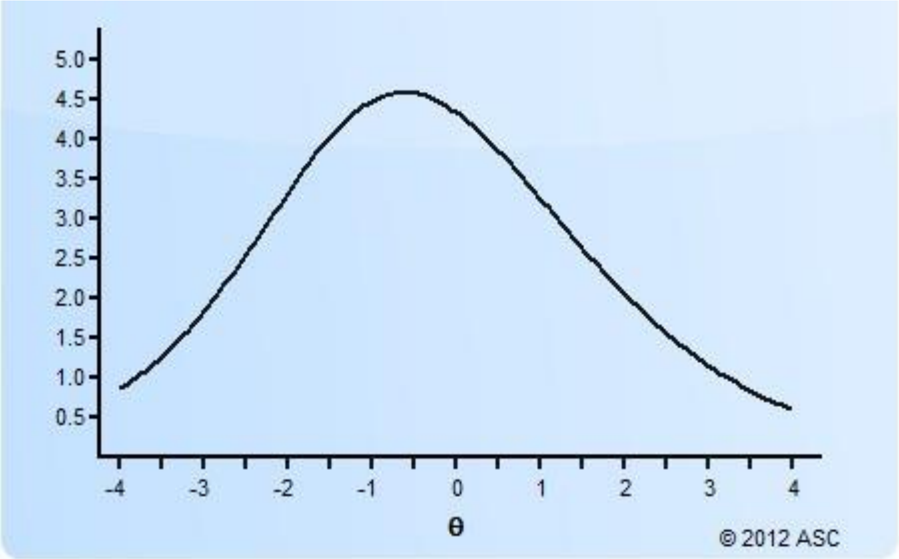


Figure 30 Test information curve for LM I Story B. The curve represents the sum of item information at various ability levels, and indicates how well the test is estimating memory over the continuum of memory scores.

LM II Story B

The serial position effect was also not evident in LM II Story B, as indicated in Figure 31. None of the regression models (linear, cubic, quadratic) fit the LM II Story B data, thus indicating that item position in the story did not clearly influence the degree to which examinees recalled a given item. In relation to immediate recall of story details, there is a less consistent pattern by which examinees recalled story details.

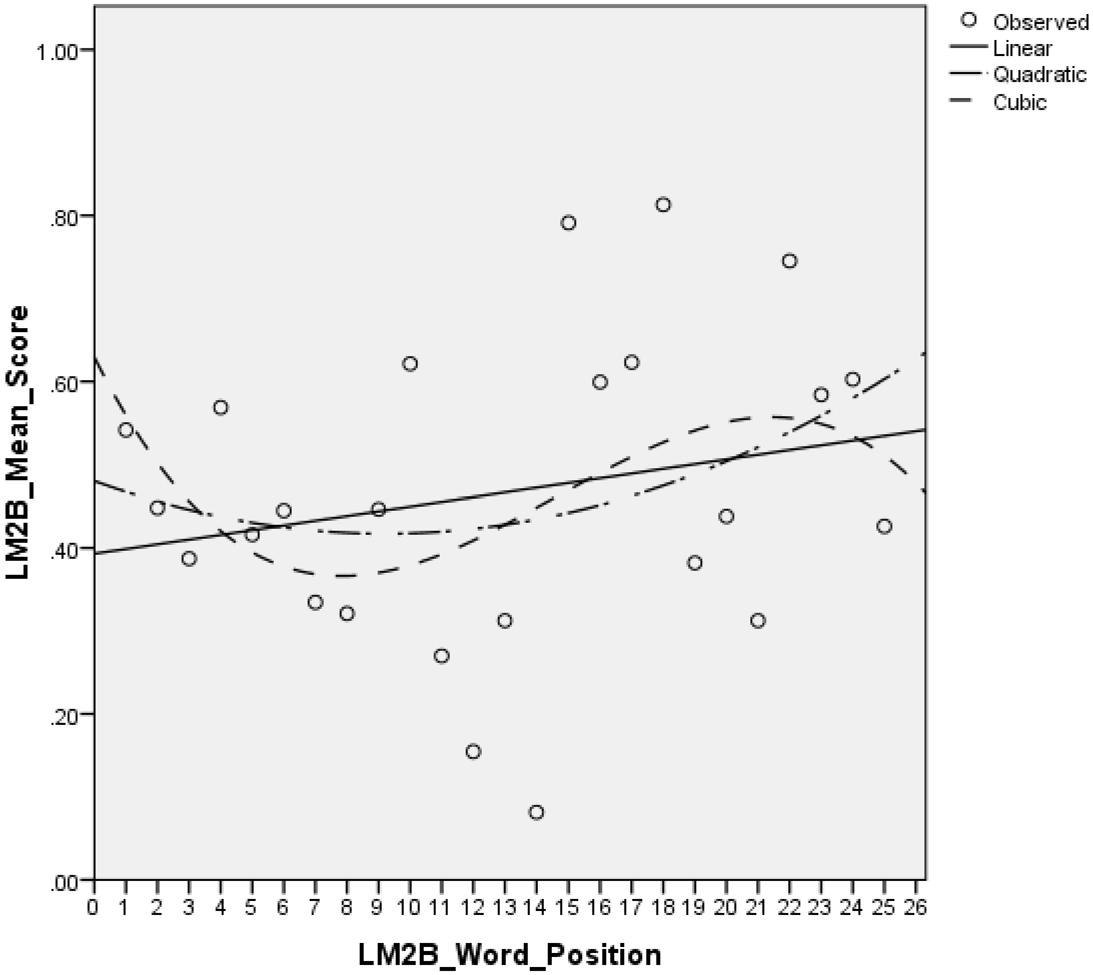


Figure 31 Regression model for LM II Story B.

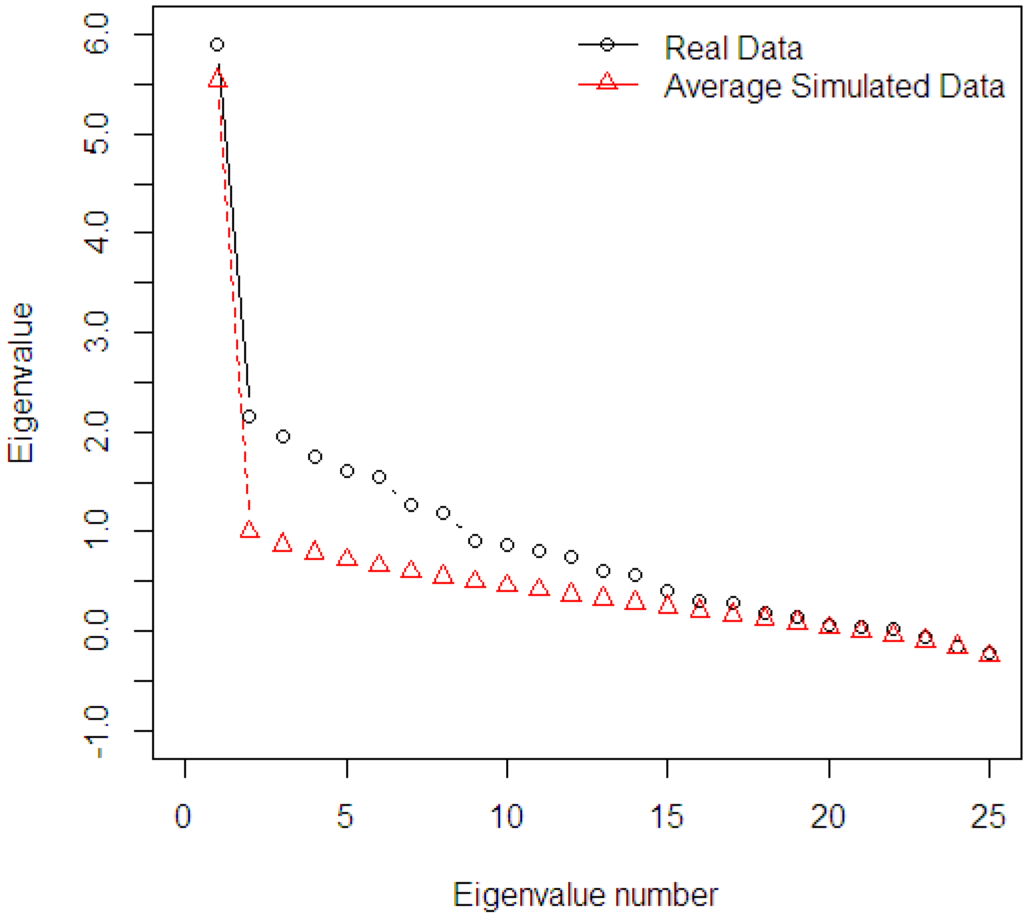


Figure 32 Scree plot for modified parallel analysis of LM II Story B.

The unidimensionality assumption was tested using modified parallel analysis. While there were again multiple eigenvalues greater than 1.00 for Story B, again, as can be observed with a visual examination of the scree plot in Figure 32, the violation of the unidimensionality assumption is not so severe as to negate the use of IRT. The proportion of examinees that answered each item correctly (P), estimates of item difficulty (β) and discrimination (α), as well as item fit statistics (Q_1 and *z*) are presented in Table 8.

Table 8 IRT item parameters, Q_1, and *z* item fit statistics for LM II Story B

Item	LM II Story B					
	P	β	α	Q_1	*p*	*z*
1	0.54	-0.22	0.47	5.37	.15	0.39
2	0.45	0.25	0.59	4.12	.25	0.43
3	0.39	0.52	0.63	3.73	.29	0.17
4	0.57	-0.36	0.48	13.95	<.01*	0.57
5	0.42	0.46	0.49	5.40	.15	0.42
6	0.45	0.31	0.47	11.46	<.01*	0.59
7	0.33	0.76	0.63	3.60	.31	0.33
8	0.32	0.98	0.50	10.34	.02*	0.38
9	0.45	0.25	0.62	6.31	.10	0.38
10	0.62	-0.73	0.42	12.10	<.01*	0.67
11	0.27	1.62	0.38	12.01	<.01*	0.69
12	0.15	2.09	0.52	3.59	.31	0.63
13	0.31	1.14	0.44	11.92	<.01*	0.42
14	0.08	2.56	0.61	36.20	<.01*	0.85
15	0.79	-1.58	0.55	14.97	<.01*	0.88
16	0.60	-0.43	0.62	6.14	.10	0.55
17	0.62	-0.46	0.77	4.81	.19	0.46
18	0.81	-1.16	1.00	5.14	.16	0.38
19	0.38	0.61	0.53	4.69	.20	0.48
20	0.44	0.28	0.66	5.07	.17	0.53
21	0.31	0.81	0.70	9.02	.03*	0.54
22	0.75	-0.91	0.90	7.48	.06	0.30
23	0.58	-0.31	0.78	17.74	<.01*	0.36
24	0.60	-0.36	0.84	5.66	.13	0.40
25	0.43	0.36	0.58	5.38	.15	0.35

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
$^*p < .05$.

Ten of the 25 items in LM II Story B did not fit the IRT model according to Q_1 analyses, but all 25 items appeared to fit the model, according to standardized residual (*z*) statistics. As in LM I Story B, several items in the middle of the story (e.g., items 10, 15, 18, and 22) appeared to have low difficulty levels. However, unlike LM I Story B, the first and last names of the main character were no longer the easiest items. Discrimination estimate was highest for Item 18 (children; α = 1.00), but in general, were quite dispersed across the 25 items (α range from 0.38 to 0.90). Item characteristic curves, item information curves, and test information curve for LM II Story B are presented in Figures 33, 34, and 35, respectively. Monotonicity was evident across all 25 items of LM II Story B. The maximum information derived from LM II Story B (5.77) is quite similar to the information derived from LM I Story B (4.59), and appears to best discriminate between examinees with similar memory ability levels (θ = -0.30 for LM II Story B, compared to θ = -0.60 for LM I Story B). Correlations among item position, item discrimination, and squared item discrimination were moderate and significant (*r* = .58, *p* < .01; r^2 = .57, *p* < .01), indicating that the assumption of local independence was violated, and item parameter estimates may be slightly biased.

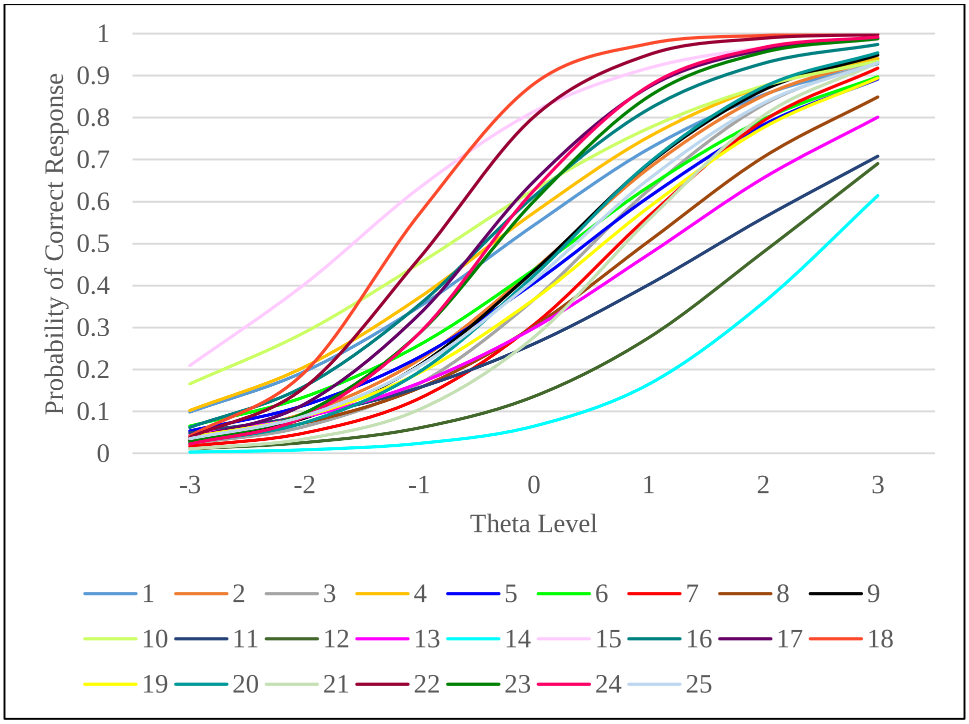


Figure 33 Item characteristic curves for LM II Story B. Each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination.

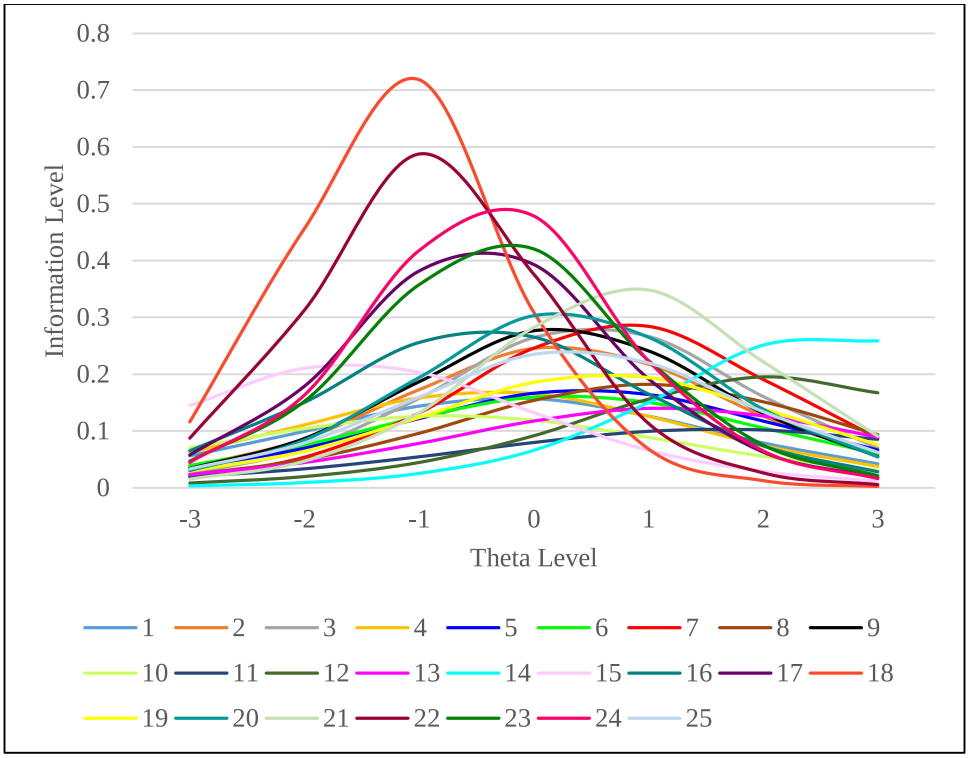


Figure 34 Item information curves for LM II Story B. Each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent.

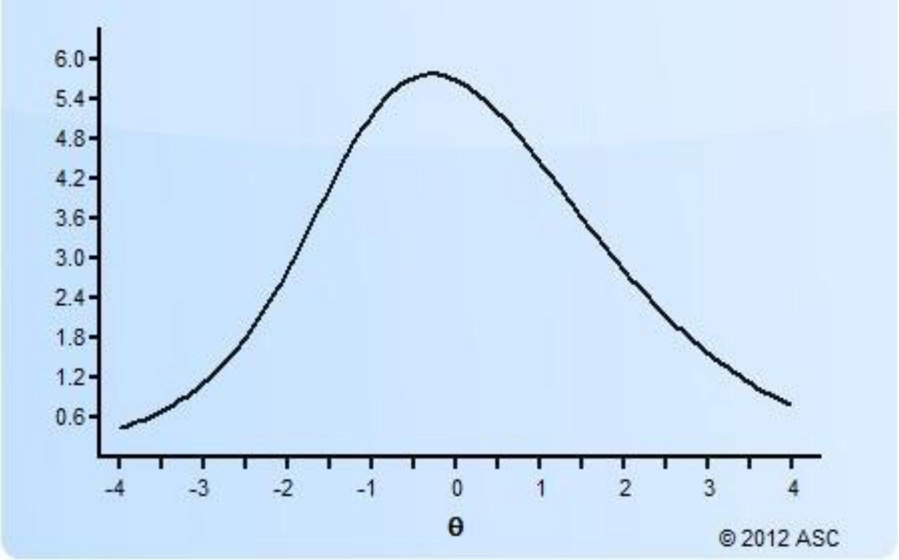


Figure 35 Test information curve for LM II Story B. The curve represents the sum of item information at various ability levels, and indicates how well the test is estimating memory over the continuum of memory scores.

LM I Story C

The serial position effect was also not apparent in LM I Story C, and none of the regression models (linear, cubic, quadratic) fit the LM I Story C data, as indicated in Figure 36. This suggests that item location in the story did not clearly influence the degree to which respondents recalled story details. However, it is notable that the first and last names of the main character were not the first few details presented in Story C, rather they are items 4 and 5.

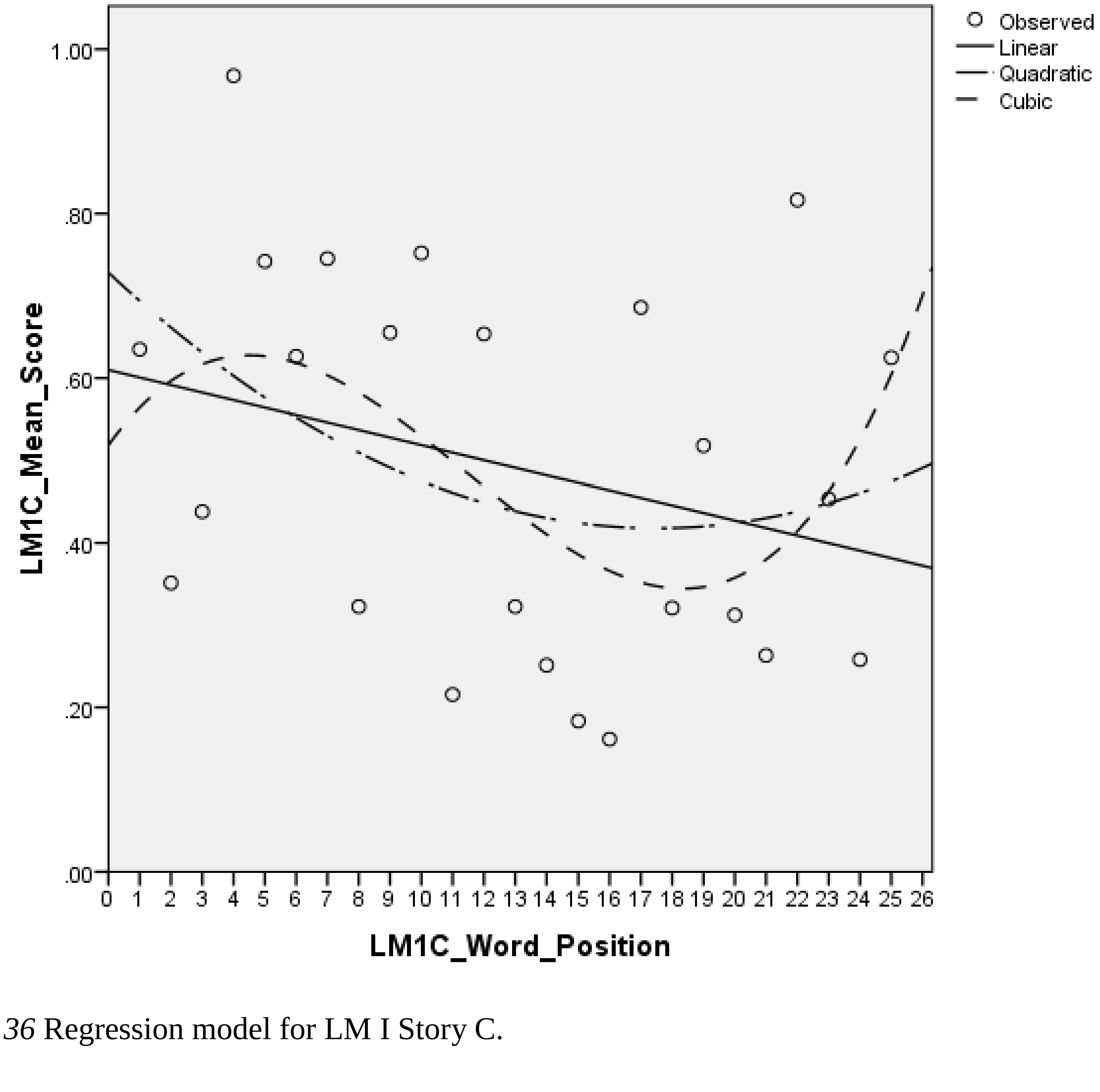


Figure 36 Regression model for LM I Story C.

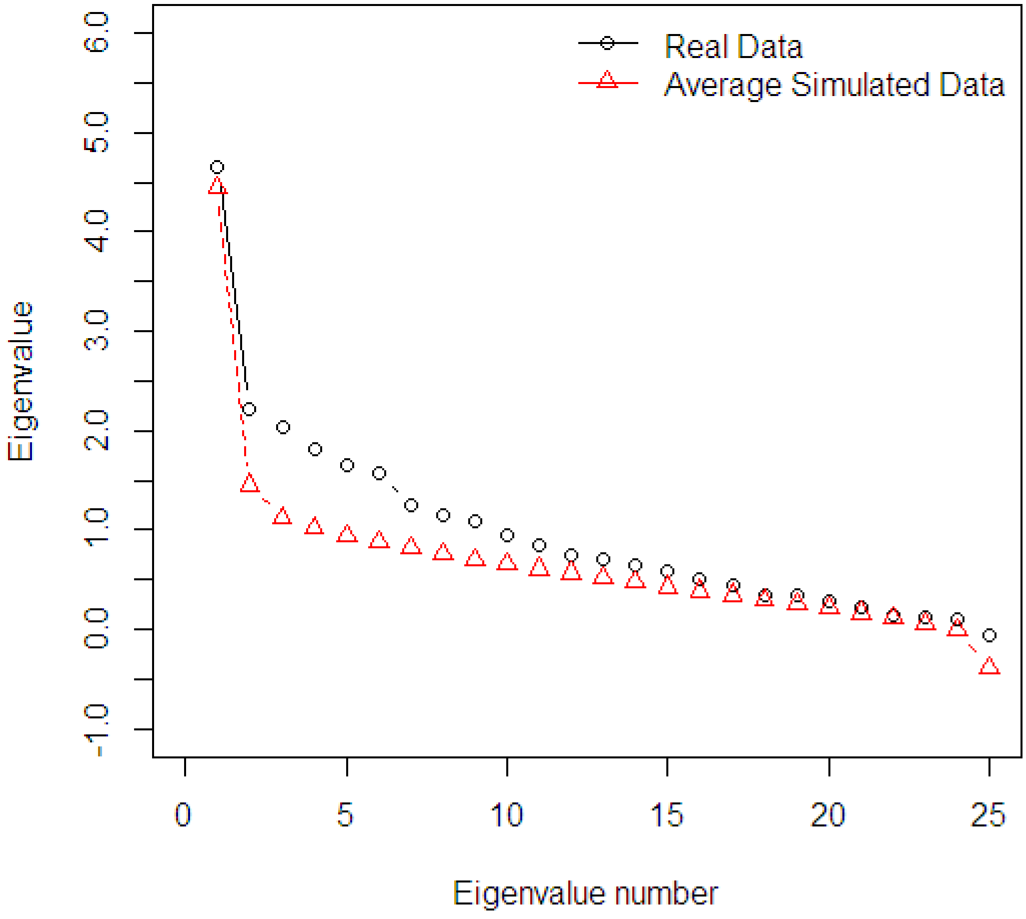


Figure 37 Scree plot for modified parallel analysis of LM I Story C.

Data did not appear unidimensional for LM I Story C, as demonstrated by the observation that there are two eigenvalues from the real data > 1.0. However, again, as can be observed with a visual examination of the scree plot in Figure 37, the violation of this assumption is not so severe as to negate the use of IRT. The proportion of examinees that answered each item correctly (P), estimates of item difficulty (β) and discrimination (α), as well as item fit statistics (Q_1 and *z*) are presented in Table 9.

Table 9 IRT item parameters, Q_1, and *z* item fit statistics for LM I Story C

Item	LM I Story C					
	P	β	α	Q_1	*p*	*z*
1	0.64	-0.64	0.57	5.81	.12	0.43
2	0.35	0.69	0.62	0.61	.89	0.17
3	0.44	0.28	0.66	1.49	.68	0.11
4	0.97	-3.12	0.71	5.50	.14	0.40
5	0.74	-1.35	0.50	6.53	.09	0.37
6	0.63	-0.87	0.36	15.56	<.01*	0.62
7	0.75	-1.97	0.32	10.49	.01*	1.03
8	0.32	1.21	0.38	3.32	.34	0.43
9	0.66	-0.78	0.53	1.89	.60	0.17
10	0.75	-1.39	0.51	9.17	.03*	0.56
11	0.22	1.68	0.49	3.11	.38	0.37
12	0.65	-0.88	0.46	7.38	.06	0.41
13	0.32	0.83	0.62	0.88	.83	0.31
14	0.25	1.09	0.71	12.79	<.01*	0.33
15	0.18	1.50	0.69	3.13	.37	0.26
16	0.16	2.22	0.47	9.98	.02*	0.80
17	0.69	-0.96	0.53	3.33	.34	0.38
18	0.32	1.01	0.48	2.62	.45	0.22
19	0.52	-0.10	0.41	1.55	.67	0.40
20	0.31	1.15	0.44	3.69	.30	0.31
21	0.26	1.39	0.48	3.79	.28	0.20
22	0.82	-1.68	0.59	7.02	.07	0.74
23	0.45	0.27	0.46	3.20	.36	0.50
24	0.26	1.73	0.37	6.18	.10	0.55
25	0.63	-0.57	0.60	3.64	.30	0.35

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
**p* < .05.

Five of the 25 items in LM I Story C did not fit the IRT model according to Q_1 analyses, but all 25 items appeared to fit the model, according to standardized residual (*z*) statistics. As was observed in LM I Story B, first (item 4) and last (item 5) names of the main character in the story were among the easiest items. In addition, items 7 (watching television), 10 (weather bulletin), 17 (rain), and 22 (stay home), arguably details making up the gist of the story, were among the items with the lowest difficulty levels.

Discrimination estimates were similar across the 25 items in LM I Story C. Item characteristic curves, item information curves, and test information curve for LM I Story C are presented in Figures 38, 39, and 40 respectively. Monotonicity was evident across all 25 items of LM I Story C, and the maximum information that can be drawn from LM I Story C is 3.77 at theta = 0.40. Parameter estimates for LM I Story C are likely unbiased, and local independence was assumed, as observed by the small and non-significant correlations between item position and item discrimination (*r* = -.18, *p* = .39; r^2 = -.20, *p* = .33).

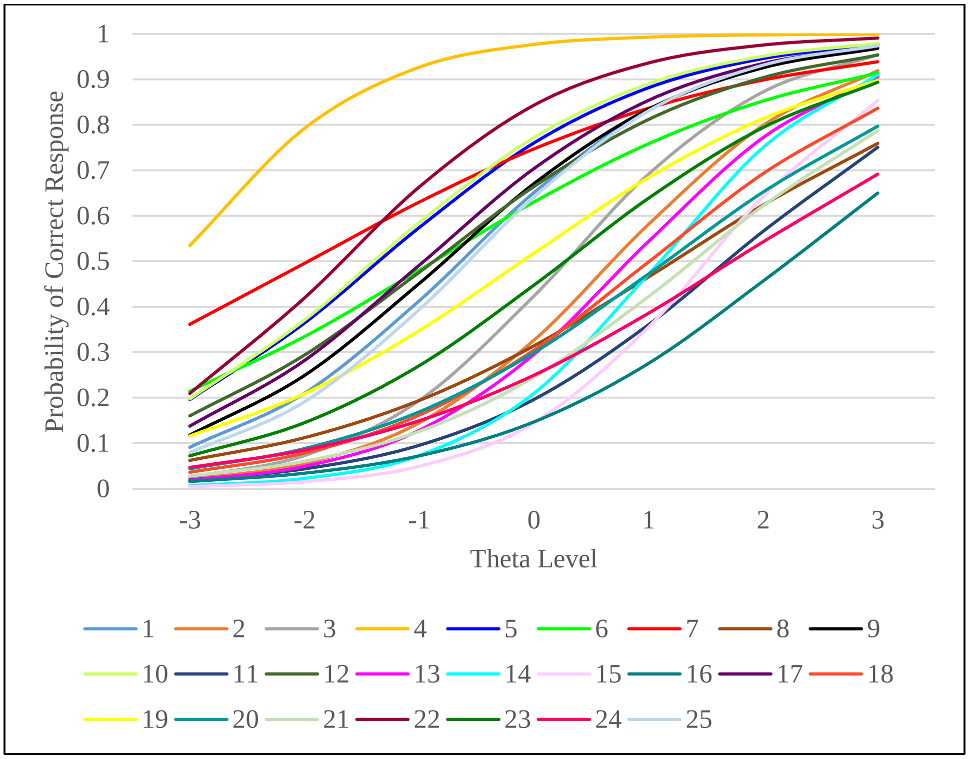


Figure 38 Item characteristic curves for LM I Story C. Each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination.

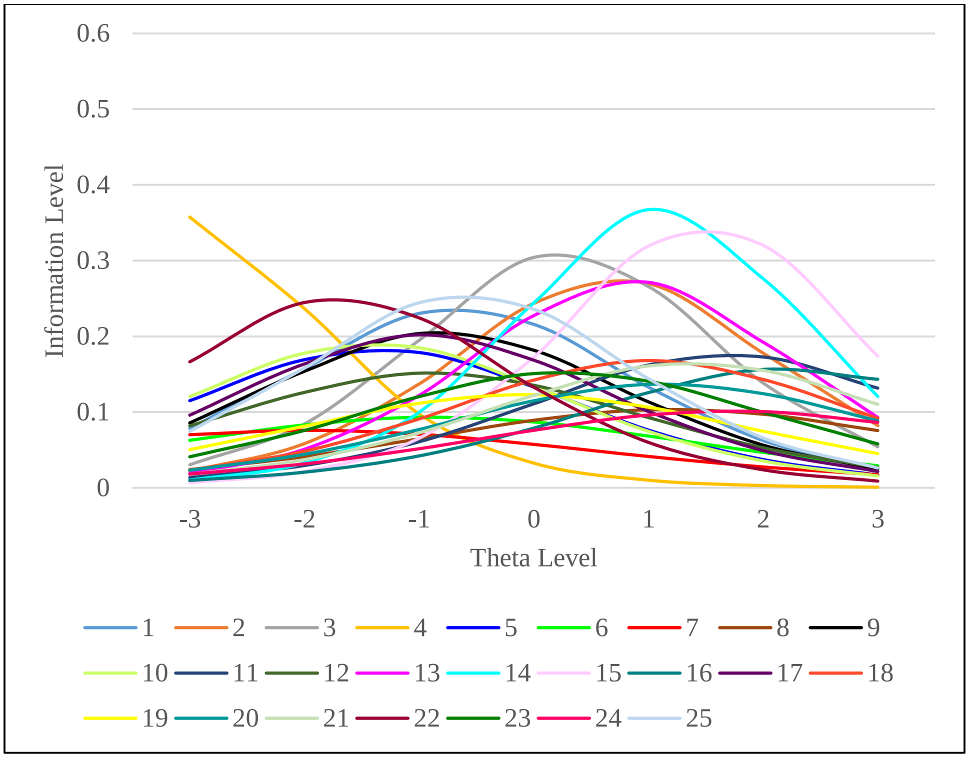


Figure 39 Item information curves for LM I Story C. Each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent.

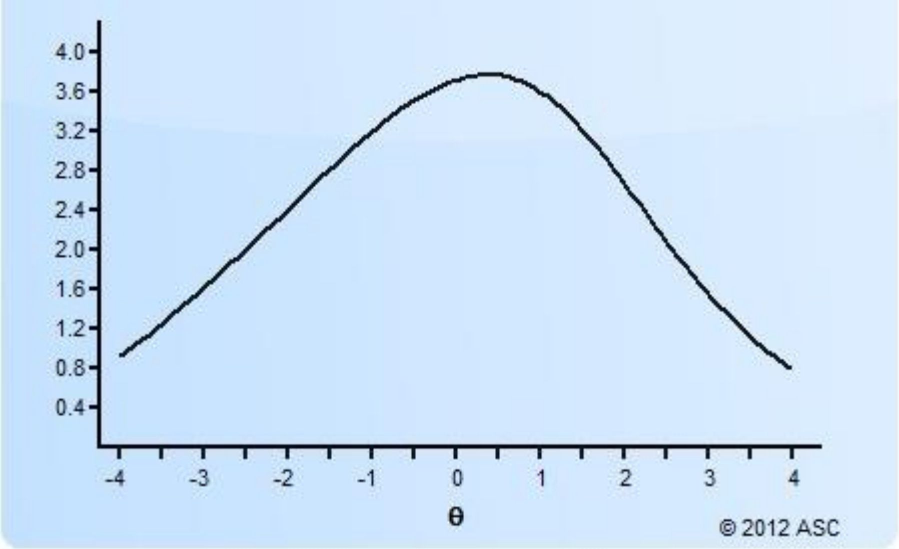


Figure 40 Test information curve for LM I Story C. The curve represents the sum of item information at various ability levels, and indicates how well the test is estimating memory over the continuum of memory scores.

LM II Story C

The serial position effect was not apparent in LM II Story C, and none of the regression models (linear, cubic, quadratic) fit the LM II Story C data, as is shown in Figure 41, once again indicating that item position in the story did not clearly influence how examinees recalled story details.

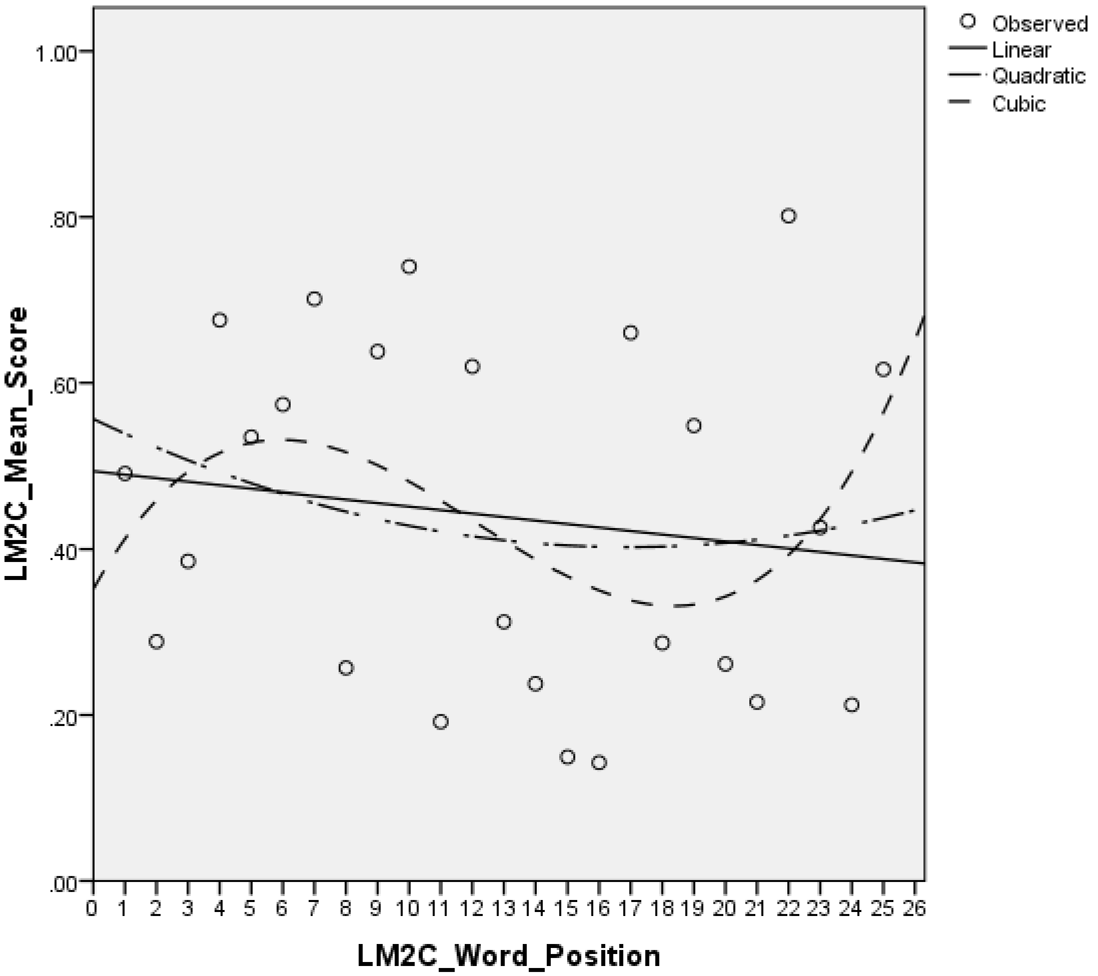


Figure 41 Regression model for LM II Story C.

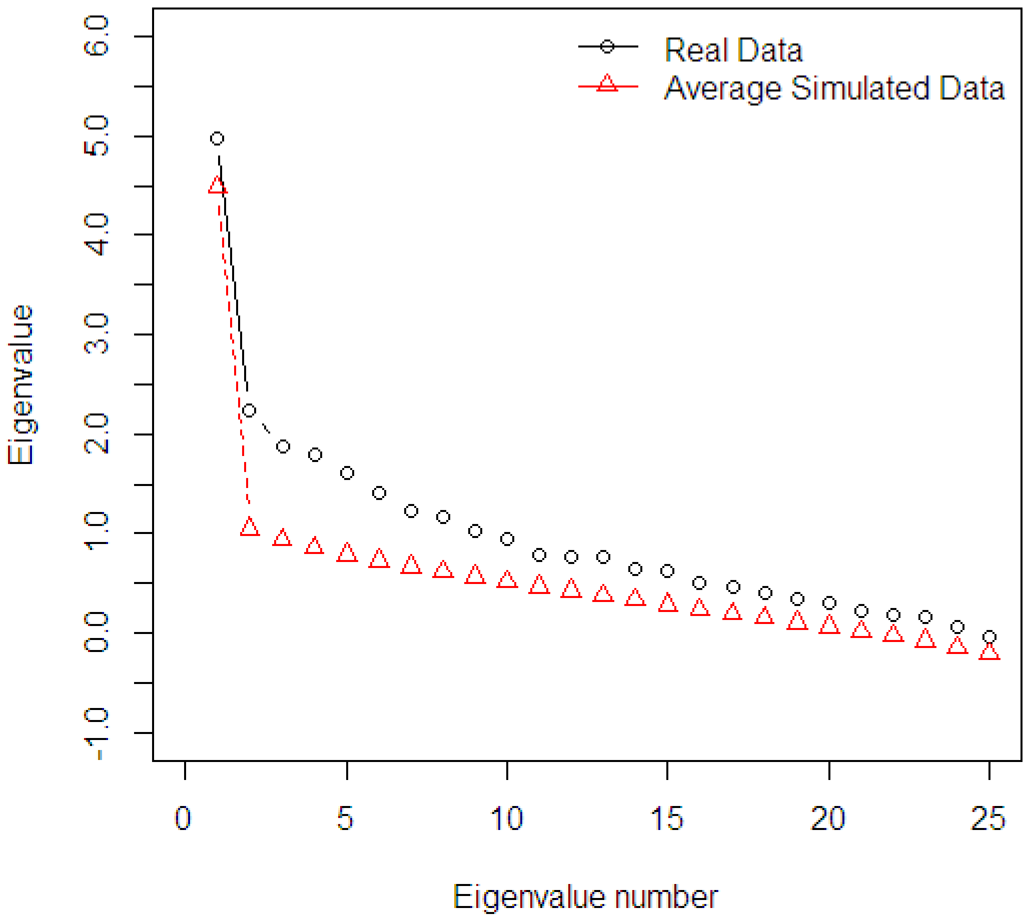


Figure 42 Scree plot for modified parallel analysis of LM II Story C.

Data did not appear unidimensional for LM II Story C, as demonstrated by the observation that there are two eigenvalues from the real data > 1.0. However, again, as can be observed with a visual examination of the scree plot in Figure 42, the violation of the unidimensionality assumption is not so severe as to negate the use of IRT. The proportion of examinees that answered each item correctly (P), estimates of item difficulty (β) and discrimination (α), as well as item fit statistics (Q_1 and *z*) are also presented in Table 10.

Table 10 IRT item parameters, Q_1, and *z* item fit statistics for LM II Story C

Item	LM II Story C					
	P	β	α	Q_1	*p*	*z*
1	0.49	0.07	0.74	13.71	<.01*	0.81
2	0.29	0.81	0.89	16.33	<.01*	0.40
3	0.39	0.42	0.92	1.54	.67	0.37
4	0.68	-0.98	0.46	7.64	.05	0.28
5	0.54	-0.15	0.52	6.81	.08	0.57
6	0.57	-0.46	0.37	14.53	<.01*	0.70
7	0.70	-1.38	0.37	11.50	<.01*	0.68
8	0.26	1.82	0.36	12.54	<.01*	0.69
9	0.64	-0.66	0.54	4.87	.18	0.37
10	0.74	-1.34	0.49	6.06	.11	0.64
11	0.19	1.77	0.53	8.27	.04*	0.33
12	0.62	-0.51	0.62	4.03	.26	0.42
13	0.31	0.89	0.63	3.04	.39	0.37
14	0.24	1.17	0.71	1.58	.66	0.32
15	0.15	2.02	0.57	6.84	.08	0.35
16	0.14	2.25	0.51	5.73	.13	0.45
17	0.66	-0.75	0.57	3.32	.34	0.70
18	0.29	1.33	0.44	7.30	.06	0.43
19	0.55	-0.33	0.33	16.44	<.01*	0.90
20	0.26	1.51	0.44	8.87	.03*	0.58
21	0.22	1.70	0.49	4.64	.20	0.43
22	0.80	-1.35	0.71	7.89	.05*	0.62
23	0.43	0.44	0.45	7.67	.05*	0.61
24	0.21	1.90	0.43	N/A	N/A	0.46
25	0.62	-0.53	0.57	N/A	N/A	0.62

Note. P = proportion of examinees who answered correctly; β = difficulty parameter; α = discrimination parameter; Q_1 = Yen's fit statistic; *p* = significance level for Q_1 statistics; *z* = Xcalibre standardized residual *z* statistic.
$^*p < .05$.

Twelve of the 25 items in LM II Story C did not fit the IRT model according to Q_1 analyses, but all 25 items appeared to fit the model, according to the standardized residual (*z*) statistics. Similar to the pattern observed in Story B Delayed Recall, the first and last names (items 4 and 5) of the main character were no longer the easiest items. However, prominent details of the story (items 7 = watching television, 10 = weather bulletin, 22 = stayed home) maintained relatively low difficulty levels. Item 3 (evening; α

= 0.92) had the highest discrimination estimate, but in general, discrimination estimates were similar across the 25 items (α range from 0.33 to 0.89). Item characteristic curves, item information curves, and test information curve for LM II Story C are presented in Figures 43, 44, and 45, respectively. Monotonicity was evident across all 25 items of LM II Story C, and the maximum information that can be drawn from LM II Story C is 4.71 at the theta level of 0.45, slightly higher than that obtained with LM I Story C (maximum information = 3.77 at theta = 0.40). Correlations among item position, item discrimination, and squared item discrimination were small and non-significant (r = -.32, p = .12; r^2 = -.37, p = .07), indicating that local independence was assumed and item parameters are likely unbiased.

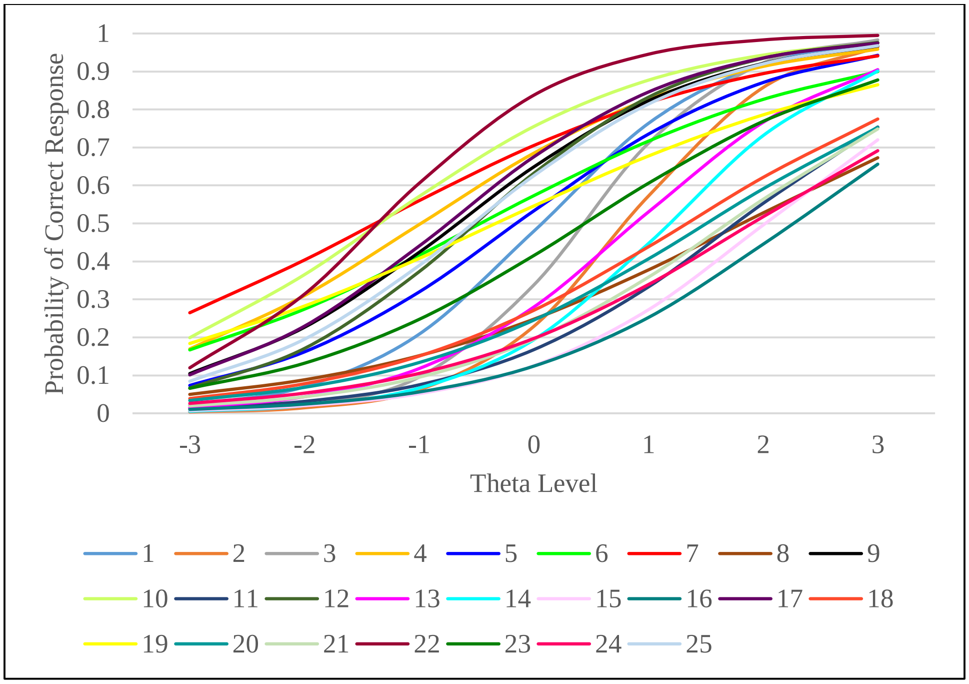


Figure 43 Item characteristic curves for LM II Story C. Each line represents the item characteristic curve for a given item, describing the probability of correct response across different trait levels. The slope of the curves at the inflection point reflects item discrimination.

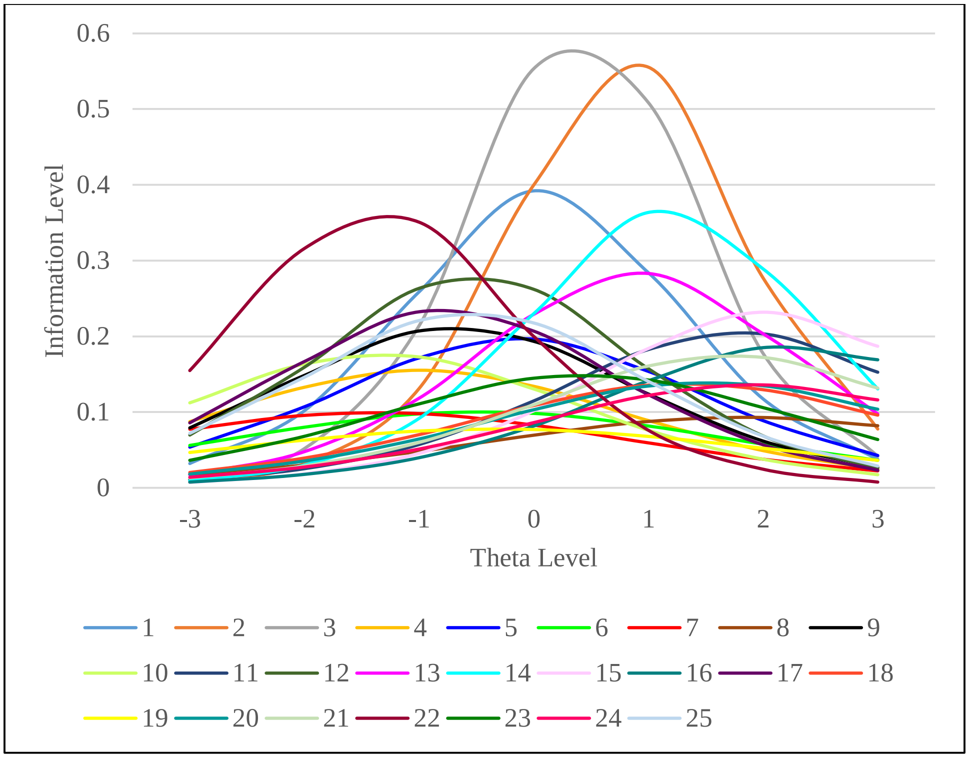


Figure 44 Item information curves for LM II Story C. Each line describes the quality of an item, based on how closely the difficulty of that item matches the ability of respondent.

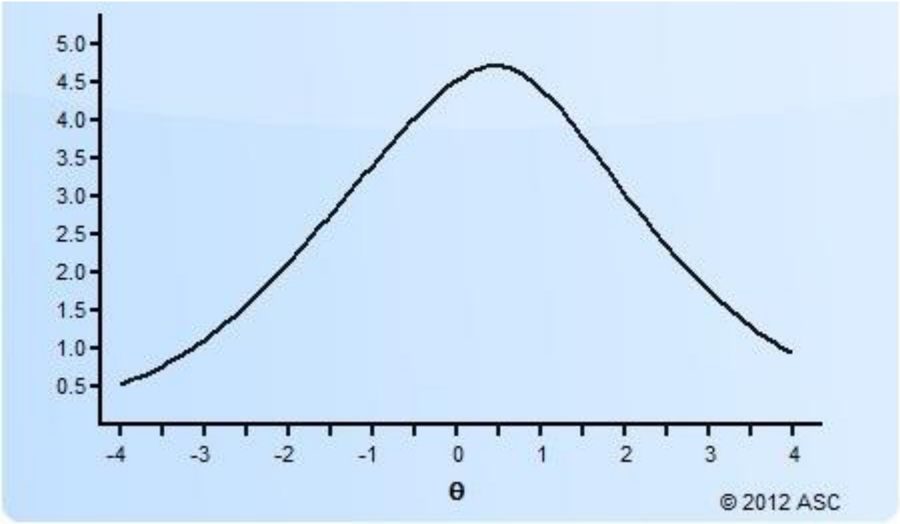


Figure 45 Test information curve for LM II Story C. The curve represents the sum of item information at various ability levels, and indicates how well the test is estimating memory over the continuum of memory scores.

Summary of IRT Results for LM trials

Overall, and as would be expected, the immediate recall trials of LM had lower mean difficulty levels (-0.21 for Story B, 0.03 for Story C), compared to the delayed recall trials (0.26 for Story B, 0.39 for Story C), and average difficulty levels between the two stories were similar (0.03 for Story B, 0.21 for Story C). In addition, the main character's first and last names were among the most readily recalled details in the immediate recall trials of Stories B and C, but this pattern was not observed in the delayed recall trials. Prominent details, or details making up the gist of the stories, were also more easily recalled across the four LM trials. Mean discrimination estimates were similar across the four trials (LM I Story B = 0.55; LM II Story B = 0.52; LM I Story C = 0.61; LM II Story C = 0.55). These findings suggest that items differ in their effectiveness in discriminating between examinees across the memory ability continuum, with prominent details more easily recalled than less prominent details.

ROC Analyses

Receiver operating characteristics (ROC) analyses were conducted to compare the conventional scoring method (assigning one point for each correct item regardless of difficulty level) with an alternative scoring method for various CVLT-II (including 14- and 16-item Trial 5) and LM trials, in differentiating the clinical sample from the research sample. Specifically, in the alternative scoring method, a weighted scoring approach is used, whereby the item with the highest difficulty level received the highest weight (16 points for CVLT-II and 25 points for LM) and the item with the lowest difficulty level received the least weight (1 point; Buschke et al., 2006). ROC curves for scores derived using the conventional scoring and weighted scoring methods were analyzed to examine the trade-off between sensitivity and specificity for each scoring algorithm for those presenting with cognitive complaints (clinical sample) compared to college students.

Results are summarized in Table 11 for CVLT-II trials. ROC graphs are presented in Appendix A. As can be seen in Table 11, there was no significant difference between ROC curves derived from the conventional scoring method and ROC curves derived from the weighted scoring approach in all trials except CVLT-II SDFR. In CVLT-II SDFR trial, ROC analyses indicated that the conventional scoring approach performed better than the weighted scoring approach in differentiating the clinical and research samples.

Table 11 ROC Analyses Results for CVLT-II Trials

	Trial 1		Trial 5		Trial 5 (14 items)		SDFR		LDFR	
	C	W	C	W	C	W	C	W	C	W
Area under curve	.66	.65	.65	.64	.64	.63	.69	.67	.66	.65
Standard error	.02	.02	.02	.02	.024	.024	.02	.02	.02	.02
95% confidence interval	.61-.69	.61-.69	.61-.69	.60-.68	.60-.69	.58-.68	.64-.73	.63-.72	.62-.71	61-.70
Positive cases	239		239		239		239			239
Negative cases	338		338		338		337			335
Difference between areas	<.01		.010		0.01		.01		0.01	
Standard error	.01		<.01		<.01		<.01		<.01	
z-statistic	.08		1.62		1.74		2.34		1.90	
Significance	.94		.11		.08		.02*		.06	

Note. C = conventional scoring; W = weighted scoring; SDFR = Short-Delay Free Recall; LDFR = Long-Delay Free Recall.
* *p* < .05, indicating a significant difference between ROC curves.

Results for ROC analyses for LM trials are presented in Table 12. Similar to the results obtained in CVLT-II trials, there was no significant difference between ROC curves derived from the conventional scoring method and ROC curves derived from the weighted scoring approach in all four LM trials. It is notable that, consistently across trials, CVLT-II trials had larger AUC values than LM trials.

Table 12 ROC Analyses Results for LM Trials

	LM I Story B		LM I Story C		LM II Story B		LM II Story C	
	C	W	C	W	C	W	C	W
Area under curve	.57	.57	.60	.59	.60	.61	.63	.63
Standard error	.02	.02	.02	.02	.02	.02	.02	.02
95% confidence interval	.52-.61	.52-.62	.55-.65	.55-.64	.55-.65	.57-.66	.59-.67	.59-.67
Positive cases	234		234		234		234	
Negative cases	355		355		355		355	
Difference between areas	<.01		<.01		.01		<.01	
Standard error	<.01		.01		.01		.01	
z-statistic	0.17		1.05		2.00		.06	
Significance	.86		.29		.05		.95	

Note. C = conventional scoring; W = weighted scoring; LM – Logical Memory.
* *p* < .05, indicating a significant difference between ROC curves.

DISCUSSION

This study utilized IRT to evaluate a common list learning task, the CVLT-II, and the WMS-IV LM subtests, which require individuals to encode and retain details from two short stories. Specifically, this study utilized IRT to identify item difficulty and discrimination parameters for these tasks. It was anticipated that item ability to discriminate between those with high and low memory functioning would vary based on the order in which items are presented to examinees. This study also investigated whether the CVLT-II or WMS-IV LM subtests more effectively quantifies verbal memory functioning. Finally, this study utilized these identified item parameters to investigate if an alternate scoring system, based on assigning more weight to items that are more difficult, could improve the precision at which verbal memory tests identify memory impairment.

Surprisingly, no known study to date has utilized IRT methods to investigate item parameters across multiple trials of a list-learning task, such as the CVLT-II. CVLT-II Trial 1 appeared to have the highest difficulty level compared to other CVLT-II trials investigated. This finding makes sense as examinees who are exposed to the list for the first time have less opportunity to develop memory strategies to facilitate deep processing relative to other trials. By the fifth and final learning trial, repeated exposures to the word list increases the likelihood that participants will recall test content. In other words, individuals with lower memory ability levels have numerous chances to improve their consolidation and recall of words so that by the fifth trial, even those with low memory ability levels have a high likelihood of recalling many words in the word list. On the other hand, because of a relatively low test ceiling, individuals with high memory ability levels do not have the same opportunity to improve performance across learning trials in the way that individuals with low memory abilities are able to do so. It is therefore not surprising that item difficulty levels in the CVLT-II were highest in Trial 1 and lower in subsequent trials. Additionally, as the item characteristic curves indicate (see Figures 4, 9, 13, 18, and 23), there was more variability in the shape and steepness of item curves within Trial 1, compared to subsequent CVLT-II trials, whereby item curves appeared more uniform. The curves indicate that with repeated trials, items functioned more similarly in terms of difficulty level, and how well they discriminate between examinees across the verbal memory ability continuum.

A different pattern of item functioning was observed with LM trials, whereby examinees were only exposed to each story once. Difficulty levels in the delayed recall conditions were similar to, or higher than, difficulty levels in the immediate recall conditions. Difficulty levels remained relatively variable within each trial, arguably related to the lack of multiple opportunities to learn the information, as well as the emotional valence or distinctiveness of a given item (for example, items 15 [robbed], 18 [children], and 22 [police] of Story B, and items 7 [watching television], 10 [weather bulletin], 17 [rain], and 22 [stay home] of Story C). Item parameters between stories B and C were comparable. Unlike the CVLT-II, where item characteristic curves became more uniform in the short- and long-delay recall trials, LM item characteristic curves remained distinct within trials.

The information previously presented allows for a comparison between measures. When comparing between CVLT-II and LM tests, item parameters obtained suggest that LM (maximum information was for θ ranging from -0.60 to 0.45) is more useful at identifying people with higher levels of θ (verbal memory ability) compared to the

CVLT-II (maximum information was for θ ranging from -1.35 to 0.10). This also makes sense, given that not a single examinee obtained a perfect score (25/25) on any of the LM trials, whereas a number of examinees obtained perfect scores (16/16) on CVLT-II Trials 5 (17% of examinees), SDFR (9% of examinees), and LDFR (11% of examinees). This finding suggests that LM has a higher score ceiling than the CVLT-II.

Although LM trials had higher difficulty levels, CVLT-II trials more strongly differentiated between examinees (as indicated by item discrimination), compared to LM trials. Specifically, with the exception of CVLT-II Trial 1 ($\alpha = 0.38$), the degree to which test items differentiated between various examinee ability levels (discrimination parameters) were higher for CVLT-II Trials 5 ($\alpha = 0.90$), SDFR ($\alpha = 0.84$) and LDFR ($\alpha = 0.89$) compared to LM trials Story B Immediate ($\alpha = 0.55$), Story B Delayed ($\alpha = 0.61$), Story C Immediate ($\alpha = 0.52$), and Story C Delayed ($\alpha = 0.55$) trials. This finding is not surprising, given that a test's discrimination parameter tends to be lower when items are measuring a broader range of ability (e.g., see Andrich, 1988) or a heterogeneous construct. As previously stated, the item characteristic curves of CVLT-II Trial 1 and of all four trials in LM were more variable, compared to CVLT-II Trials 5, SDFR, and LDFR. CVLT-II trials (except Trial 1) had higher discrimination values because items functioned in a similar way and at similar difficulty levels, which allows a more precise identification of subtle differences in memory ability levels.

Given these findings, LM may be more aptly administered if a clinician is evaluating memory functioning in someone expected to be within the average range of ability, whereas the CVLT-II may be more aptly administered to someone with suspected memory impairments, as it more precisely discriminates between examinees with lower ability levels. Nevertheless, it is important to note that the CVLT-II and LM tests differ in terms of type of memory measures (less structured information in the form of a word list vs. structured information in the form of a story). Therefore, in many clinical situations it seems appropriate to administer both tests, as they provide different types of information, as well as different levels of discrimination at different verbal memory ability levels.

An important question to address, then, is what influences item parameters within trials? In other words, what makes one item easier than another? One of the main hypothesis for explaining differences in the degree to which items differentiate between examinees of varying ability levels and the degree to which items are readily remembered is the serial position effect. The research literature suggests that the immediate recall of a list of words is likely influenced by two interrelated cognitive abilities, which has been postulated to be attention (recency items) and short-term episodic memory (primacy and middle items; Gavett & Horwitz, 2011; Buschke et al., 2006).

In the present study, as expected, the serial position effect was observed in CVLT-II Trial 1 and to a lesser extent, in CVLT-II Trial 5. However, this effect was not observed in CVLT-II SDFR and CVLT-II LDFR, or any of the four LM trials. One possible explanation for this finding in the CVLT-II trials is that examinees had the opportunity, after 5 learning trials and a short delay, to effectively consolidate the words in memory, for example, by remembering words according to categories. Serial position effects are typically observed across different word lists presented and recalled once, rather than the same list recalled five times (Geffen, Moar, O'Hanlon, Clark, & Geffen, 1990).

Generally, experimental memory research conducted with neurologically intact subjects indicate that primacy recall is associated with long-term storage, whereas recency recall is associated with short-term storage (Glanzer & Cunitz, 1966). When a list is presented in the same order for multiple trials, as is done with the CVLT-II, there can be more variability in whether responses are being generated from short-term or long-term storage (Massman, Delis, & Butters, 1993). Therefore, it is not surprising that the serial-position effect is less prominent in Trial 5, and not observed in CVLT-II SDFR and LDFR. The finding of a lack of observed serial position effect for SDFR and LDFR is consistent with research findings, given the brief delay period and interference of List B for SDFR, as well as the 20 minute delay and various nonverbal interference tasks for LDFR. Specifically, researchers have demonstrated that in free recall, the recency effect is attenuated after a delay (Howard & Kahana, 1999), and almost eliminated after 15 seconds of a distractor task (Glanzer & Cunitz, 1966; Postman & Phillips, 1965).

The lack of serial position effect in the LM subtests in this study is not entirely consistent with previous research findings. For example, Hall and Bornstein (1991) found that the serial position effect was evident in the immediate recall of a brief story, both among patients with head injuries, and normal controls. That study made use of Wechsler Memory Scale-Revised (WMS-R; 1987) Logical Memory Story A, which is very similar in terms of length and emotional content to Story B in this research. A plausible explanation for the observed discrepancy in findings between this study and those of Hall and Bornstein (1991) is the methodology used to evaluate the serial position effect. Hall and Bornstein defined primacy items as the first eight items, middle items as the next nine items, and recency items as the final eight items in their story recall task. They then used repeated measures ANOVA and found that there was a significant difference between the percentage of items recalled for each third of the story, among both patients with head injury and normal controls. Their study evaluated only the immediate recall trial of one story (Story A of the WMS-R LM subtest). When utilizing a repeated measures ANOVA approach (as opposed to quadratic regression approach), results from the present study quite closely match the results by Hall and Bornstein for the research participants, but not the clinical sample, and only in Story B, which is similar to the WMS-R LM Story A that Hall and Bornstein used. Specifically, in the present study, research participants recalled a significantly higher percentage of primacy and recency items than middle items (F [2, 662.64] = 50.18, *p* < .01) in the immediate recall of Story B. However, the clinical sample did not demonstrate a significant difference in memory of primacy, recency, and middle items (F [2, 466] = 0.96, *p* = .33). On the other hand, for the immediate recall of Story C, both the research and clinical samples did not demonstrate a significant difference in percentage of items recalled between middle and recency items (mean difference between middle and recency items for research sample = 0.01, *p* = .37; mean difference between middle and recency items for clinical sample = 0.02, *p* = .20). Descriptive statistics and ANOVA results for the LM trials are presented in Table 13 below. Although some differences are observed in the degree to which primacy, middle, and recency items are recalled, the differences are not consistently observed across different samples and different stories. This suggests that scoring items based simply on item position may not improve measurement precision. Mean differences observed are likely related to emotional or distinctive items located at various positions in the stories, as opposed to true primacy and recency effects.

Table 13 Mean (and SD) Percentage words recalled in first, middle, and last section of stories

	Primacy (8 Items)	Middle (9 items)	Recency (8 items)	F
LM Story B Immediate				
Research sample	62.32 (21.32)	48.92 (18.41)[a]	58.31 (25.41)[a, b]	50.18*
Clinical sample	54.22 (23.45)	46.11 (21.42)[a]	55.98 (24.60)[b]	17.72
LM Story B Delayed				
Research sample	48.42 (24.83)	45.16 (19.11)	55.07 (26.22)[a, b]	23.92*
Clinical sample	35.47 (24.86)	40.55 (20.50)[a]	51.87 (26.68)[a, b]	48.97*
LM Story C Immediate				
Research sample	64.30 (19.37)	44.82 (20.06)[a]	45.99 (20.34)[a]	141.29*
Clinical sample	54.38 (19.23)	40.55 (20.50)[a]	42.47 (22.99)[a]	51.37
LM Story C Delayed				
Research sample	54.03 (21.35)	43.35 (18.75)[a]	44.28 (19.87)[a]	39.91*
Clinical sample	40.97 (23.09)	37.46 (22.06)[a]	38.78 (21.88)	2.66

Note. [a] Significantly different recall in comparison to primacy items. [b] Significantly different recall in comparison to middle items.

In another study evaluating the serial position effect in the recall of stories, Messier, Gagnon and Knott (1997) found that male, but not female, adults remembered more primacy items (7 items) in a story, compared to middle (9 items) and recency (7 items) items. It appears that although the serial position effect can be observed in some instances of story recall, this effect is not as robust as some may consider, and it is not consistently observed across samples.

The presence of the serial position effect in some CVLT-II trials and different levels of distinctiveness and emotional valence in LM trials contribute to different difficulty and discrimination parameters in different items and trials. Therefore, it does not make intuitive sense to assign the same weights to both easy and difficult items. Previous studies have proposed a weighted scoring approach to assign higher weights to items that are more difficult. For example, as previously mentioned, Buschke and colleagues (2006) assigned weights to items ranging from 1 to 10 based on item position in the immediate recall trial of a 10-item word list and found that the weighted scoring system improved test precision (increasing ROC AUC from 0.77 to 0.86, $p < .05$), and increased the effect size difference in scores between those with mild Alzheimer's Disease and those without dementia from 1.08 to 1.52. The present study assigned weights to items based on item difficulty parameters provided by IRT analyses. Contrary to expectations, ROC analyses utilizing a weighted scoring approach based on obtained difficulty parameters indicated few improvements in discriminating between a clinical and research sample compared to the conventional scoring approach (see Tables 11 and 12).

A number of differences between the present study and the research by Buschke et al. (2006) could explain why the current weighted scoring approach did not result in improvements in test discrimination. First, the study by Buschke and colleagues involved a 10-item word-list consisting of semantically unrelated words, administered over the phone, and words from the list were repeated if requested by the examinee. In contrast, the present study utilized a 16-item word list consisting of semantically related words, administered in-person, and repetition of words from the list within a given administration trial was not allowed. It is possible that differences in word list length and administration methods affected the degree to which results from the two studies converged, given that the probability of recalling words decreases as list length increases (Unsworth & Engle, 2006), and repetition of words help improve recall performance. In addition, the serial position effect is expected to be more prominent in CVLT-II Trial 1, given the long word list, but diminishes over the trials as examinees adopt different encoding and retrieval strategies (i.e., recalling words in order vs. recalling words according to semantic categories).

Second, Buschke and colleagues (2006) compared those with Alzheimer's disease (diagnosed at consensus case conferences by professionals who were blind to word list results), with non-dementia controls, whereas the present study utilized a more heterogeneous clinical sample (simply those who presented at a neuropsychology clinic with cognitive complaints). The groups utilized in ROC analyses by Buschke and colleagues were more easily distinguishable compared to the groups utilized in this study. Specifically, the raw score effect size between their clinical and control sample was extremely large (Cohen's *d* = 1.08), compared to the effect size of 0.57 for CVLT-II Trial 1 in this study. Additionally, the pattern of performance differences across items differed dramatically between studies. Buschke and colleagues (2006) control group displayed the serial position curve, whereas their clinical group did not (see Figure 46). In contrast the current study included clinical and nonclinical groups that performed similarly (see Figure 47).

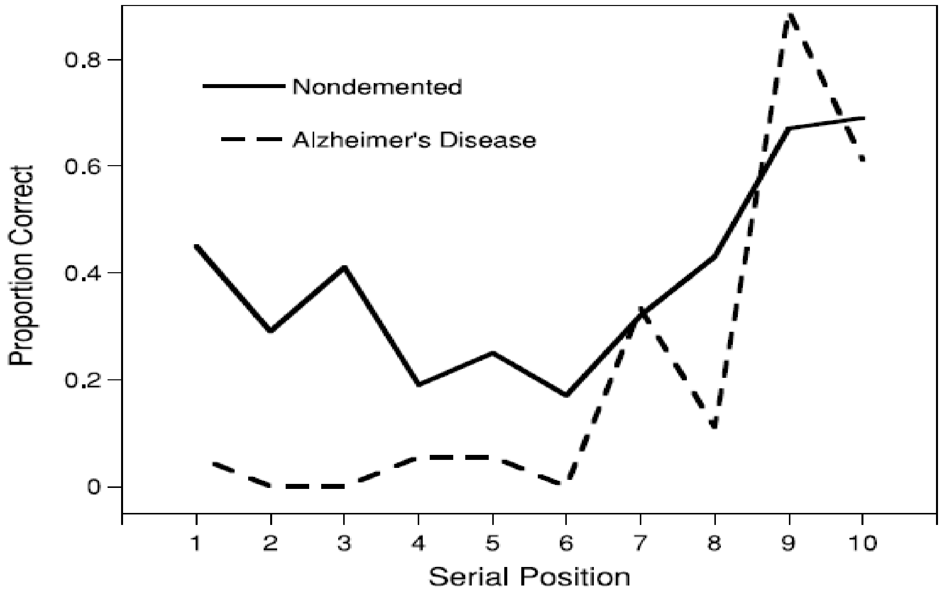


Figure 46 Serial position curves for immediate free recall by older adults with and without Alzheimer’s disease in Buschke et al (2006).

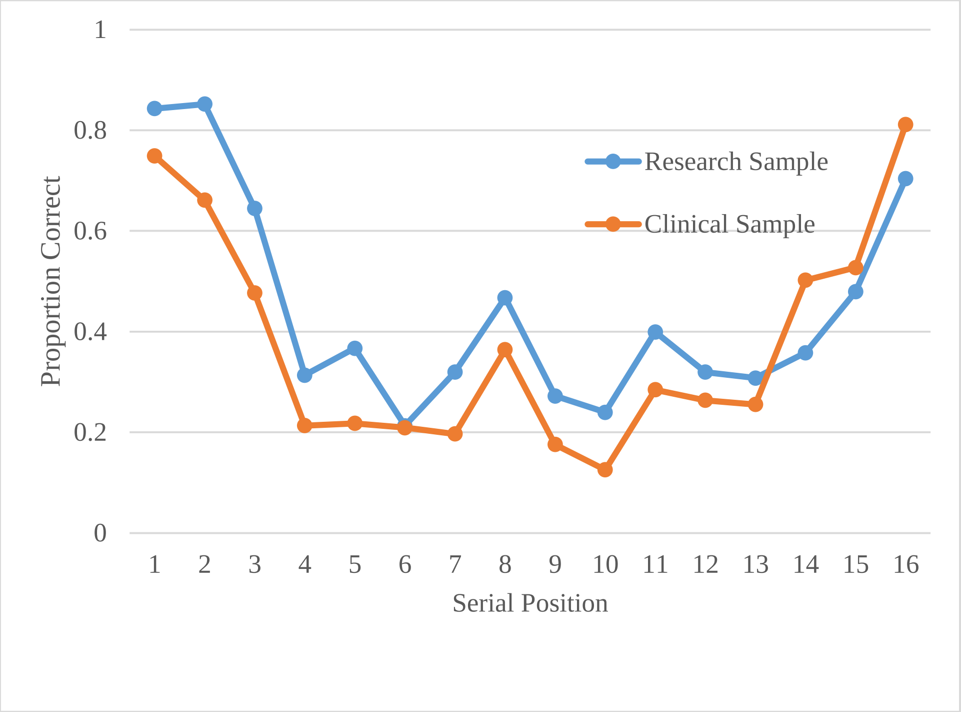


Figure 47 Serial position curves for free recall by research and clinical samples.

In the present study, although there is a higher likelihood that patients at the neuropsychology clinic had memory difficulties compared to college students, it is also likely that there were outpatients from the neuropsychology clinic presenting with other difficulties that may not necessarily result in poor memory, such as depression, ADHD and LD. Similarly, although most college students in a research pool do not have memory difficulties, there is a possibility that some college students may have had impaired short-term and long-term memory. This potential overlap in functioning between groups is likely to impact evaluation of weighted scoring approaches by confounding ROC analyses.

To address this, several exploratory ROC analyses were conducted using other criterion/categorization variables to indicate memory impairment, such a CVLT-II LDFR z-score of less than -1.4 when evaluating LM trials (impaired group n = 60; intact group n = 348), and LM II scaled scores below 6 in evaluating CVLT-II Trials (impaired group n = 41; intact group n range[1] = 367 to 370). When using a criteria of CVLT-II LDFR z-score of less than -1.4, the effect size differences in LM raw scores between impaired and intact groups ranged from 0.60 (LM I Story B) to 0.95 (LM II Story B), and were still somewhat smaller than the large group difference observed in Buschke at al.'s (2006) study. ROC analyses comparing the weighted and conventional scoring of LM trials indicated that the weighted scoring approach was not significantly better than the conventional scoring approach in differentiating between impaired and non-impaired performances on CVLT-II LDFR. Results from these ROC analyses are presented in Table 14. Compared to the ROC analyses using clinical vs. research samples presented

[1] CVLT T1 and T5 were based on N = 411; SDFR based on n = 410; LDFR based on n = 408

earlier (see Table 12), AUCs in these analyses were larger, which makes sense, given the more extreme differentiation in performance between groups.

Table 14 ROC analyses results for LM trials using LDFR z-scores as comparison groups

	LM I Story B		LM I Story C		LM II Story B		LM II Story C	
	C	W	C	W	C	W	C	W
Area under curve	.66	.64	.70	.69	.75	.75	.72	.71
Standard error	.04	.04	.04	.04	.04	.04	.03	.04
95% confidence interval	.61 - .71	.59 - .69	.65 - .74	.64 - .74	.70 - .79	.70 - .79	.67 - .76	.67 - .76
Positive cases	60		60		60		60	
Negative cases	348		348		348		348	
Difference between areas	.02		.01		<.01		.01	
Standard error	.01		.01		.01		.01	
z-statistic	1.30		.65		.01		.52	
Significance	.19		.52		.99		.60	

Note. C = conventional scoring; W = weighted scoring; LM = Logical Memory.
* *p* < .05, indicating a significant difference between ROC curves.

When using a criteria of LM II SS of less than 6 in evaluating CVLT-II trials, the effect size differences in CVLT-II scores ranged from 1.17 (CVLT-II Trial 5) to 1.37 (CVLT-II SDFR) between impaired and intact groups, which is larger than the effect size of score difference between Buschke and colleagues' (2006) clinical and control samples. Even so, ROC analyses comparing weighted and conventional scoring of CVLT-II trials indicated that the weighted scoring approach was still not significantly more effective than the conventional scoring approach in differentiating between impaired and non-impaired performances on LM II. Results of these ROC analyses are presented in Table 15. Once again, when compared to the CVLT-II ROC analyses using clinical vs. research samples presented earlier (see Table 11), AUCs in these analyses were larger, which again, is not surprising, given the more differentiated comparison groups. Therefore, similarities between weighted and conventional scoring AUCs in these analyses indicate that differences in the memory ability levels of the clinical vs. research samples cannot fully account for the difference in results obtained between the present study and those obtained by Buschke and colleagues.

Table 15 ROC analyses results for CVLT-II trials using LM II scaled scores as comparison groups

	Trial 1		Trial 5		Trial 5 (14 items)		SDFR		LDFR	
	C	W	C	W	C	W	C	W	C	W
Area under curve	.83	.79	.80	.79	.80	.79	.83	.83	.83	.81
Standard error	.03	.04	.04	.04	.04	.03	.04	.04	.04	.04
95% confidence interval	.79 - .86	.74 - .82	.75 - .83	.74 - .83	.76 - .84	.75 - .83	.79 - .87	.79 - .87	.79 - .86	.77 - .85
Positive cases	41		41		41		41		41	
Negative cases	370		370		370		369		367	
Difference between areas	.04		.01		.01		<.01		.02	
Standard error	.02		.01		.01		.01		.01	
z-statistic	1.99		0.89		0.96		0.33		1.57	
Sig.	.05		.38		.34		.74		.12	

Note. C = conventional scoring; W = weighted scoring; SDFR – Short-Delay Free Recall; LDFR – Long-Delay Free Recall; Sig. = significance.
* *p* < .05, indicating a significant difference between ROC curves.

An additional factor to consider when trying to account for differences in findings between this study and those of Buschke et al. (2006) is the manner in which scoring weights were assigned. Buschke and colleagues assigned weights solely based on item position on the list, whereas this study assigned scoring weights based on item difficulty parameters. Perhaps alternative weighted scoring approaches could be employed, given that a somewhat arbitrary approach was utilized in assigning scoring weights (scoring weights were assigned within the range of 1 to 16 for the CVLT-II trials, and 1 to 25 for the LM trials). In an effort to briefly explore other scoring approaches, several alternative weighted scoring methods were used, investigating CVLT-II Trial 1 scores, given that the first trial of the list-learning test was used in Buschke et al's study and because the first trial appears to have the most diverse item difficulty parameters of the CVLT-II trials investigated. First, scores were assigned solely based on item position (as was done in the study by Buschke and colleagues) so that Item 1 obtained a score of 16, and Item 16 obtained a score of 1. Utilizing this scoring approach, and consistent with the findings of Buschke et al., the weighted scoring approach (AUC = 0.65) was significantly more precise than the conventional scoring approach (AUC = 0.62) in differentiating the clinical sample from the research sample (*p* = .03). However, when this weighted scoring approach was used to differentiate memory impairment as measured by a LM II Scaled Score of < 6, the two scoring approaches functioned in a similar manner (weighted scoring AUC = 0.82; conventional scoring AUC = 0.83; *p* = .54).

A second scoring alternative involved assigning CVLT-II Trial 1 items with difficulty (β) levels ≥ 0 with a score of 1 for a correct response and items with difficulty (β) levels < 0 with a score of 0 for both a correct and incorrect response. The logic behind this weighting scheme is that difficulty levels < 0 involve examinees with low memory ability levels answering the item correctly. Theoretically, such items do not appear to discriminate effectively between impaired and intact memory. ROC analyses revealed that there was no significant difference between these two scoring methods when differentiating between research and clinical samples (weighted scoring AUC = 0.60;

conventional scoring AUC = 0.62; *p* = .35). However, in contrast to expectation, the conventional scoring approach (AUC = 0.83) performed significantly better than this weighted scoring approach (0.75) when differentiating memory impairment as measured by a LM II Scaled Score of < 6 (*p* < .01).

Finally, a third scoring alternative involved assigning CVLT-II Trial 1 items with difficultly (β) levels ≥ 0 with a score of 2 for a correct response and items with difficulty (β) levels < 0 with a score of 1 for a correct response. The logic behind this approach was to investigate whether simply assigning a higher score to items that are more difficult, without making score discrepancies too large by assigning a large range of weights could help differentiate between impaired and intact memory. In addition, assigning a score of 1 for easy items acknowledges accurate recall of items, no matter how easy that recall. This third weighted scoring approach also did not significantly differ from the conventional scoring approach, both in terms of differentiating the clinical and research samples (weighted scoring AUC = .62; conventional scoring AUC = .62; *p* = .85), or identifying memory impairment as defined by a LM II Scaled Score < 6 (weighted scoring AUC = .81; conventional scoring AUC = .83; *p* = .15). In summary, it is quite apparent that attempts at differentially weighting specific items to improve prediction of memory impairment did not significantly improve upon the conventional scoring approach of simply assigning one point to each item.

It is important to note, however, that this study did not exhaustively explore all potential alternative weighted scoring approaches. It is commonly agreed upon that a standard scoring method based simply on the sum of items answered correctly is theoretically suboptimal. For example, suppose a test has 20 easy items with extremely small variance and another 10 moderately difficult items with a large variance. The 20 items are likely adding a constant to the 10 item score, and contribute little to the variance of the composite score (Rudner, 2001). Simply using a sum-of-scores approach would be suboptimal as the 20 items would weigh twice the amount of the 10 items, even though the 20 items provide little information about the examinee. Given this, continued efforts to understand how CVLT-II and LM item parameters might be utilized to improve test precision are warranted. Future studies could more exhaustively investigate other weighted scoring approaches and other CVLT-II and LM trials to determine if the precision of these tests could be improved. Numerous other weighted scoring methods could be explored. Rudner (2001) highlighted other possible methods of weighted scoring, including assigning greater weights to items with higher reliability or validity.

Similarly, other trials and indicators of both the CVLT-II and LM could be studied more extensively to determine if they might provide more information about verbal memory ability levels. For example, the total score for each word across the five immediate recall trials may provide greater discrimination levels, given that it incorporates more information about patients, compared to a score for each word obtained in only one trial. Alternatively, recognition trials may provide more information about patients' ability to encode verbal material, and pinpoint weaknesses in information retrieval instead. More closely evaluating the item parameters of items in the recognition portions of memory test may also provide useful information about memory functioning.

In addition, future studies could compare item parameters of different word lists (e.g., CVLT-II, Hopkins Verbal Learning Test [HVLT; Brandt, 1991], Rey Auditory Verbal Learning Test [RAVLT; Rey, 1964]) to determine if semantically related word lists affect the learning curve differently than semantically unrelated word lists, or to determine how differences in number of learning trials affect the degree to which words at various positions in the word list are recalled. Similarly, future studies could also investigate different story memory tests (e.g., LM of WMS-IV compared to Story Memory of RBANS [Randolph, 1998]) to evaluate how story length, distinctive details, and emotional valence differentiate between examinees of different memory ability levels.

A possible clinical limitation to this study is that the approach taken was to individually evaluate each trial to see if different assigned weights could predict impaired memory. This relatively narrow focus on aspects of a test is not typically undertaken in clinical settings. For example, CVLT-II immediate learning and retention is typically evaluated based on a total score, summed across five trials, and LM scores are evaluated based on the sum of scores across Stories B and C. Anecdotally, clinicians typically include information about Total Trial 1-5 scores, SDFR, SD Cued Recall, LDFR, and LD Cued Recall in their clinical reports. Performances on individual learning trials tend to be summed into a total score, given the variability in performance observed across each trial. Performance on individual immediate recall trials, let alone individual items, are rarely looked at for interpretation, given the smaller reliability of one trial or individual item, as compared to a scale comprising many trials and many items. Within a managed care setting, where there is limited time available to meet with patients and make sense of clinical data, it may not be practical for clinicians to evaluate each individual trial. On the other hand, if one specific trial was found to be extremely effective in quantifying memory functioning (for example, item parameters between stories B and C were comparable and performed similarly in terms of differentiating between a clinical and research sample, as well as in terms of identifying memory impairment in CVLT-II LDFR), a clinician could save time and resources by administering only a portion of a longer test.

While not directly related to the primary research questions, to further explore the previous issue raised, this project allows clinicians to consider whether it is necessary to administer multiple list learning trials, or two stories when quantifying memory functioning. For example, given the similarity of LM Stories B and C item parameters, additional exploratory ROC analyses were conducted to investigate if Stories B and C function similarly in discriminating between groups defined by CVLT-II LDFR z-score less than -1.4. Two comparisons were initially conducted: (1) LM I Story B vs LM I Story C, (2) LM II Story B vs LM II Story C. The total recall scores for LM I Story B (AUC = 0.66) and LM I Story C (AUC = 0.70) did not significantly differ in terms of how well memory impairment was identified (*p* = .21). LM II Story B (AUC = 0.75) and LM II Story C (AUC = 0.72) also functioned similarly in identifying memory impairment on CVLT-II LDFR (*p* = .34). These results suggest that the stories function equivalently, and that it might not be necessary to administer both. Given this, next, it was evaluated whether Delayed Recall predicted group membership more effectively than Immediate Recall (i.e., LM I Story B vs LM II Story B, and LM I Story C vs. LM II Story C). Results indicate that LM II Story B (AUC = .75) was significantly more effective in identifying memory impairment on LDFR compared to LM I Story B (AUC = .66; *p* < .01). However, LM II Story C (AUC = .72) did not significantly differ from LM I Story C (AUC = .70) in terms of effectiveness of identifying memory impairment on LDFR (*p* =

.30). Therefore, future studies could further investigate the potential tradeoffs associated with administering *just* LM I Story C to quantify memory functioning in research participants and patients referred for neuropsychological evaluations.

Another limitation associated with this study is that there were a number of violations in the assumptions underlying IRT. For example, across trials in the CVLT-II and LM tests, modified parallel analyses consistently revealed that data was not clearly unidimensional. The lack of unidimensionality of CVLT-II Trial 1 is consistent with the finding of Gavett & Horwitz (2011), who also found that RAVLT Trial 1 was bi-dimensional. In addition, there were some violations of the assumption of local independence, as indicated by the finding of a significant correlation between item discrimination and item position for CVLT-II Trial 1, and LM B Immediate and Delayed Trials. Although such violations of unidimensionality and local independence are common and invariably found in other IRT studies (see Childs et al., 2000; Gavett & Horwitz, 2011; Thomas & Locke, 2010 for examples), it is nevertheless important to acknowledge that findings from this study may not be as robust across different samples. As such, additional research should be conducted across other samples to investigate the replicability and generalizability of the present findings.

Despite these limitations, this study clearly demonstrated that IRT is a useful method to further understand test psychometric properties. Specifically, this study demonstrated that the serial position effect was evident in CVLT-II Trials 1 and 5, but not in the CVLT-II SDFR, CVLT-II LDFR, and any of the LM trials. This study also found that multiple weighted scoring approaches based on IRT-derived item difficulty levels did not seem to improve test precision. Finally and most importantly, this study documented the item parameters in the trials evaluated and demonstrated that the CVLT-II provided, on average, greater discrimination of memory abilities, compared to LM, but that LM had higher difficulty levels. In conclusion, this study highlights that it is important to critically consider the item properties and scoring approaches undertaken in day-to-day clinical activities to quantify memory functioning. Findings from this study should improve clinical practice with the recognition that some items or trials perform better than others in discriminating between examinees with low levels of memory ability, and that it is important to more closely evaluate item properties of tests used in clinical decision-making. Identification of test item properties is likely to increase the precision and diagnostic utility of these measures, and more generally clinical assessment, which will improve patient outcomes.

REFERENCES

Alexander, M. P., Stuss, D. T., & Fansabedian, N. (2003). California Verbal Learning Test: Performance by patients with focal frontal and non-frontal lesions. *Brain, 126,* 1493-1503.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (AERA; 1999). *Standard for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

American Psychological Association Division 40. (2012). Division 40: Clinical Neuropsychology. Retrieved from http://www.div40.org/

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, *1*(4), 363-378.

Ankenmann, R. (1994). Goodness of fit and ability estimation in the graded response model. Unpublished manuscript. In Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24(1), 50-64.

Atkinson, R. C., & Shiffrin, R. M. (1965). Mathematical models for memory and learning. Technical report no. 79. Institute for mathematical studies in the social sciences. Stanford University, CA.

Baddeley, A. D. (2002). The psychology of memory. In A. D. Baddeley, M. D. Kopelman, & B. A. Wilson (Eds.), The Handbook of Memory Disorders, Second Edition (pp. 3-15). West Sussex, England: John Wiley & Sons Ltd.

Baker, F. N., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques.* Second Edition. New York, NY: Marcel Dekker.

Baldo, J. V., Delis, D., Kramer, J., & Shimamura, A. P. (2002). Memory performance on the California Verbal Learning Test-II: Findings from patients with focal frontal lesions. *Journal of the International Neuropsychological Society, 8,* 539-546.

Benton, A. L. (1994). Neuropsychological assessment. *Annual Review of Psychology, 45,* 1-23.

Benton, A. L., & Hamsher, K. de S. (1976*). Multilingual Aphasia Examination*. Iowa City: University of Iowa. In Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology, 17,* 380-392.

Benton, A. L., Sivan, A., Hamsher, K., Varney, N., & Spreen, O. (1994). *Contributions to neuropsychology assessment: A clinical manual* (2nd ed.). New York: Oxford University Press.

Bianchini, K. J., Mathias, C. W., & Greve, K. W. (2001). Symptom validity testing: A critical review. *The Clinical Neuropsychologist, 15*(1), 19-45.

Bigler, E. D., Johnson, S. C., Anderson, C. V., Blatter, D. D., Gale, S. D., … & Abildskov, T. J. (1996). Traumatic brain injury and memory: The role of hippocampal atrophy. *Neuropsychology, 10,* 333-342.

Bigler, E. D., Rosa, L., Schultz, F., Hall, S., & Harris, J. (1989). Rey-Auditory Verbal Learning and Rey-Osterrieth Complex Figure Design performance in Alzheimer's disease and closed head injury. *Journal of clinical psychology,45*(2), 277-280.

Blessed G. T., Roth B. E., & Tomlinson, M. (1968). The association between quantitative measures of dementia and of senile changes in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry, 114,* 797-811. In Mungas, D., & Reed, B. R. (2010). Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in Medicine, 19,* 1631-1644.

Bondi, M. W., Monsch, A. U., Galasko, D., Butters, N., Salmon, D. P., & Delis, D. C. (1994). Preclinical cognitive markers of dementia of the Alzheimer type. *Neuropsychology, 8*(3), 374.

Brandt, J. (1991). The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist, 5*(2), 125-142.

Brown, L. B., & Storandt, M. (2000). Sensitivity of category cued recall to very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology, 15,* 529-534.

Buschke, H., Sliwinski, M. J., Kuslansky, G., Katz, M., Verghese, J., & Lipton, R. B. (2006). Retention weighted recall improves discrimination of Alzheimer’s disease. *Journal of the International Neuropsychological Society, 12,* 436-440.

Calamia, M., Markon, K., Denburg, N. L., & Tranel, D. (2011). Developing a short form of Benton’s Judgment of Line Orientation Test: An Item Response Theory Approach. *The Clinical Neuropsychologist, 25,* 670-684.

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review, 13,* 181-197.

Chen, S. H. A., Kareken, D. A., Fastenau, P. S., Trexler, L. E., & Hutchins, G. D. (2003). A study of persistent post-concussion symptoms in mild head trauma using positron emission tomography. *Journal of Neurology, Neurosurgery & Psychiatry, 74*(3), 326-332.

Childs, R. A., Dahlstrom, W. G., Kemp, S. M., & Panter, A. T. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 Depression scale. *Assessment, 7,* 37-54.

Claparede, E. (1951). Recognition and me-ness. Translated in D. Rapaport (Ed.), *Organization and Pathology of thought* (pp. 58-75). New York: Colombia University Press. (Originally published, 1911).

Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment: An introduction to tests and measurement.* New York, NY: McGraw-Hill.

Corkin, S. (1968). Acquisition of motor skill after bilateral medial temporal-lobe excision. *Neuropsychologia*, *6*(3), 255-265.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., … & van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology, 61,* 1018-1027.

Cronbach, L. J. (1984). *Essentials of psychological testing.* New York, NY: Harper & Row.

Crosson, B., Novack, T. A., Trenerry, M. R., & Craig, P. L. (1988). California Verbal Learning Test (CVLT) performance in severely head-injured and neurologically normal adult males. *Journal of Clinical and Experimental Neuropsychology*, *10*(6), 754-768.

Cullum, C. M., Butters, N., Troster, A. I., & Salmon, D. P. (1990). Normal aging and forgetting rates on the Wechsler Memory Scale-Revised. *Archives of Clinical Neuropsychology, 5,* 23-30.

de Jager, C. A., Hogervorst, E., Combrinck, M., & Budge, M. M. (2003). Sensitivity and specificity of neuropsychological tests for mild cognitive impairment, vascular cognitive impairment and Alzheimer's disease.*Psychological medicine, 33*(06), 1039-1050.

Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology, 54,* 180-187.

DeJong, J., & Donders, J. (2009). A confirmatory factor analysis of the California Verbal Learning Test – Second Edition (CVLT-II) in a traumatic brain injury sample. *Assessment, 16,* 328-336.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *California Verbal Learning Test–Second Editi*on. San Antonio, TX: Psychological Corporation.

DeVellis, R. F. (2006). Classical test theory. *Medical care, 44*(11), S50-S59.

Donders, J. (2008). A confirmatory factor analysis of the California Verbal Learning Test – Second Edition (CVLT-II) in the standardization sample. *Assessment, 15,* 123-131.

Donnell, A. J., Belanger, H. G., & Vanderploeg, R. D. (2011). Implications of psychometric measurement for neuropsychological interpretation. *The Clinical Neuropsychologist, 25,* 10097-1118.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95,* 134-135.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology, Vol. I* (pp. 577-636). Palo Alto, CA: Consulting Psychologists Press.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Emilien, G., Antoniadis, E., Durlach, C., Maloteaux, J.-M., & van der Linden, M. (2004). *Memory: Neuropsychological, imaging, and psychopharmacological perspectives.* New York, NY: Psychology Press.

Faraggi, D., & Reiser, B. (2002). Estimation of the area under the ROC curve.*Statistics in medicine*, *21*(20), 3093-3106.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). “Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189-198.

Flaugher, R. L. (1978). The many definitions of test bias. *American Psychologist*, 671-679.

Frazier, T. W., Naugle, R. I., & Haggerty, K. A. (2006). Psychometric adequacy and comparability of the short and full forms of the Personality Assessment Inventory. *Psychological Assessment, 18,* 324-333.

Garfield, S. L. (1974). *Clinical psychology: The study of personality and behavior.* USA: Aldine Transaction.

Gavett, B. E., & Horwitz, J. E. (2012). Immediate list recall as a measure of short-term episodic memory: Insights from the serial position effect and item response theory. *Archives of Clinical Neuropsychology, 27,* 125-135.

Gazzaniga, M., Ivry, R. B., & Mangun, G. R. (2008). *Cognitive Neuroscience: The Biology of the Mind.* Third Edition. New York, NY; W. W. Norton & Company, Inc.

Geffen, G., Moar, K. J., O'hanlon, A. P., Clark, C. R., & Geffen, L. B. (1990). Performance measures of 16–to 86-year-old males and females on the auditory verbal learning test. *The Clinical Neuropsychologist*, *4*(1), 45-63.

Gibbons, L. E., Crane, P. K., Mehta, K. M., Pedraza, O., Tang, Y., Manly, J. J., ... & Mungas, D. (2011). Multiple, correlated covariates associated with differential item functioning (DIF): Accounting for language DIF when education levels differ across languages. *Ageing Research*, *2*(1), 19.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall.*Journal of verbal learning and verbal behavior*, *5*(4), 351-360.

Gold, J. M., Hermann, B. P., Randolph, C., Wyler, A. R., Goldberg, T. E., & Weinberger, D. R. (1994). Schizophrenia and temporal lobe epilepsy: a neuropsychological analysis. *Archives of general psychiatry, 51*(4), 265-272.

Grzybowski, M., & Younger, J. G. (1997). Statistical methodology: III. Receiver operating characteristic (ROC) curves. *Academic Emergency Medicine*, *4*(8), 818-826.

Gulliksen, H. (1950). *Theory of mental tests.* New York, NY: Wiley.

Guyer, R., & Thompson, N.A. (2012). *User's Manual for Xcalibre item response theory calibration software, version 4.1.8*. St. Paul MN: Assessment Systems Corporation.

Hall, S., & Bornstein, R. A. (1991). Serial-position effects in paragraph recall following mild closed-head injury. *Perceptual and Motor Skills, 72,* 1295-1298.

Hall, K. S., Gao, S., Emsley, C. L., Ogunniyi, A. O., Morgan, O., Hendrie, H. C. (2000). Community screening interview for dementia (CSI ''D''); performance in five disparate study sites. *Int J Geriatr Psychiatry,15, 521-531.*

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications, Inc.

Harvill, L. M. (1991). An NCME instructional module on standard error of measurement. *Instructional Topics in Educational Measurement,* 33-41.

Helmstaedter, C., Wietzke, J., & Lutz, M. T. (2009). Unique and shared validity of the "Wechsler logical memory test", the "California verbal learning test", and the "verbal learning and memory test" in patients with epilepsy. *Epilepsy research, 87*(2), 203-212.

Herlitz, A., Nilsson, L. G., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & cognition*, *25*(6), 801-811.

Herlitz, A., & Rehnman, J. (2008). Sex differences in episodic memory. *Current Directions in Psychological Science, 17*(1), 52-56.

Hoelzle, J. B., Nelson, N. W., & Smith, C. A. (2011). Comparison of Wechsler Memory Scale–Fourth Edition (WMS–IV) and Third Edition (WMS–III) dimensional structures: Improved ability to evaluate auditory and visual constructs. *Journal of Clinical and Experimental Neuropsychology, 33*(3), 283-291.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577-601.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied psychological measurement*, *6*(3), 249-260.

Humphreys, L. G., & Montanelli Jr, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10*(2), 193-205.

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Immekus, J. C., & Maller, S. J. (2009). Item parameter invariance of the Kaufman Adolescent and Adult Intelligence Test across male and female samples. *Educational and Psychological Measurement, 69,* 994-1012.

Jacobs, M. L., & Donders, J. (2007). Criterion validity of the California Verbal Learning Test-(CVLT-II) after traumatic brain injury. *Archives of clinical neuropsychology*, *22*(2), 143-149.

Kaitaro, T., Koskinen, S., & Kaipio, M. L. (1995). Neuropsychological problems in everyday life: A 5-year follow-up study of young severely closed-head-injured patients. *Brain Injury, 9,* 713-727.

Kent, G. P., Schefft, B. K., Howe, S. R., Szaflarski, J. P., Yeh, H.-S., & Privitera, M. D. (2006). The effects of duration of intractable epilepsy on memory function. *Epilepsy and Behavior, 9,* 469-477.

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words?.*Memory & cognition*, *31*(8), 1169-1180.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I.* New York, NY: Academic Press.

La Femina, F., Senese, V. P., Grossi, D., & Venuti, P. (2009). A battery for the assessment of visuo-spatial abilities involved in drawing tasks. *The Clinical Neuropsychologist*, *23,* 691-714.

Lavenex, P., & Amaral, D. G. (2000). Hippocampal-neocortical interaction: A hierarchy of associativity. *Hippocampus*, *10*(4), 420-430.

Less, J. M. (2014). PEIP: Functions for Aster Book on inverse theory. Retrieved from the World Wide Web: http://cran.r-project.org/web/packages/PEIP/index.html

Lees-Haley, P. R., Smith, H. H., Williams, C. W., & Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical survey. *Archives of Clinical Neuropsychology, 11,* 45-51.

Levy, D. A., Manns, J. R., Hopkins, R. O., Gold, J. J., Broadbent, N. J., & Squire, L. R. (2003). Impaired visual and odor recognition memory span in patients with hippocampal lesions. *Learning & Memory*, *10*(6), 531-536.

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favoring women in verbal but not in visuospatial episodic memory. *Neuropsychology*, *15*(2), 165.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). The practice of neuropsychological assessment. In M. D. Lezak, D. B. Howieson, & D. W. Loring. *Neuropsychological assessment* (pp. 3-14). New York, NY: Oxford University Press.

Lindhiem, O., Kolko, D. J., & Yu, L. (2013). Quantifying diagnostic uncertainty using item response theory: The Posterior Probability of Diagnosis Index. *Psychological Assessment, 25,* 456-466.

Linn, R. T., Wolf, P A., Bachman, D. L., Knoefel, J. E., Cobb, J. L., … & D’Agostino, R. B. (1995). The ‘preclinical phase’ of probably Alzheimer’s Disease: A 13-year prospective study of the Framingham Cohort. *Archives of Neurology, 52*, 485-490.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Weasley.

Lutz, J. T. (2012). *Item response theory and factor analysis applied to the Neuropsychological Symptom Scale (NSS).* Retrieved from bsu.edu.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*, 321-943.

Makatura, T. J., Lam, C. S., Leahy, B. J., Castillo, M. T., & Kalpakjian, C. Z. (1999). Standardized memory tests and the appraisal of everyday memory. *Brain Injury, 13,* 355-367.

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61,* 793-817.

Massman, P. J., Delis, D. C., & Butters, N. (1993). Does impaired primacy recall equal impaired long-term storage?: Serial position effects in Huntington's disease and Alzheimer's disease. *Developmental Neuropsychology*, *9*(1), 1-15.

Massman, P. J., Delis, D. C., Butters, N., Dupont, R. M., & Gillin, J. C. (1992). The subcortical dysfunction hypothesis of memory deficits in depression: neuropsychological validation in a subgroup of patients. *Journal of Clinical and Experimental Neuropsychology*, *14*(5), 687-706.

Mattis, S. (1988). *Dementia Rating Scale*. Odessa, FL: Psychological Assessment Resources. In Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology, 17,* 380-392.

MedCalc Statistical Software version 14.12.0 (2014). MedCalc Software bvba, Ostend, Belgium; http://www.medcalc.org.

Messier, C., Gagnon, M., & Knott, V. (1997). Effect of glucose and peripheral glucose regulation on memory in the elderly. *Neurobiology of Aging, 18*, 297-304.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R.,… & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues.

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

Milner, B. (1966). Amnesia following operation on the temporal lobes. In C.W.M. Whitty & O. L. Zangwill (eds), Amnesia. London: Butterworth. In Baddeley, A. D. (2002). The psychology of memory. In A. D. Baddeley, M. D. Kopelman, & B. A. Wilson (Eds.), The Handbook of Memory Disorders, Second Edition (pp. 3-15). West Sussex, England: John Wiley & Sons Ltd.

Milner, B. (1971a). Disorders of learning and memory after temporal lobe lesions in man. *Clinical neurosurgery, 19,* 421-446.

Milner, B. (1971b). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin, 27,* 272-277.

Mokros, A., Schilling, F., Eher, R., & Nitschke, J. (2012). The Severe Sexual Sadism Scale: Cross-validation and scale properties. *Psychological Assessment*, *24*(3), 764.

Moore, B. A., & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury*, *18*(10), 975-984.

Mungas, D., & Reed, B. R. (2010). Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in Medicine, 19,* 1631-1644.

Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & Gonzalez, H. (2004). Spanish and English Neuropsychological Assessment Scales: Further development and psychometric characteristics. *Neuropsychology, 16,* 347-359.

Mungas, D., Reed, B. R., Haan, M. N., & Gonzalez, H. (2005). Spanish and English Neuropsychological Assessment Scales: Relationship to demographics, language, cognition, and independent function. *Neuropsychology, 19,* 466-475.

Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology, 17,* 380-392.

Mungas, D., Reed, B. R., Marshall, S. C., & Gonzalez, H. M. (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. *Neuropsychology, 14,* 209-223.

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., Van Belle, G., Fillenbaum, G., et al. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology,
39, 1159–1165. In Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology, 17,* 380-392.

Nordlund, A., Rolstad, S., Klang, O., Lind, K., Hansen, S., & Wallin, A. (2007). Cognitive profiles of mild cognitive impairment with and without vascular disease. *Neuropsychology, 6,* 706-712.

Nordlund, A., Rolstad, S., Klang, O., Lind, K., Pedersen, M., Blennow, K., … & Wallin, A. (2008). Episodic memory and speed/attention deficits are associated with Alzheimer-typical CSF abnormalities in MCI. *Journal of the International Neuropsychological Society, 14,* 582-590.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

Pedraza, O., & Mungas, D. (2008). Measurement in cross-cultural neuropsychology. *Neuropsychology Review*, *18*(3), 184-193.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly journal of experimental psychology*, *17*(2), 132-138.

Piotrowski, C., & Zalewski, C. (1993). Training in psychodiagnostic testing in APA-approved PsyD and PhD clinical psychology programs. *Journal of Personality Assessment, 61,* 394-405.

Qualls, C., Bliwise, N., & Stringer, A. (2000). Short forms of the Benton Judgment of Line Orientation test: Development and psychometric properties. *Archives of Clinical* Neuropsychology, 15, 159-163.

R Core Team (2014). The R project for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

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.

Rabin, L. A., Paré, N., Saykin, A. J., Brown, M. J., Wishart, H. A., Flashman, L. A., & Santulli, R. B. (2009). Differential memory test sensitivity for diagnosing amnestic mild cognitive impairment and predicting conversion to Alzheimer's disease. *Aging, Neuropsychology, and Cognition, 16*(3), 357-376.

Randolph, C. (1998). *RBANS manual: Repeatable battery for the assessment of neuropsychological status*. San Antonio, TX: The Psychological Corporation.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127-137.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*(1), 45-58.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8,* 164-184.

Rey, A. (1964) L'Examen Clinique en Psychologie.Paris: Press Universitaire de France. In Gavett, B. E., & Horwitz, J. E. (2012). Immediate list recall as a measure of short-term episodic memory: Insights from the serial position effect and item response theory. *Archives of Clinical Neuropsychology, 27,* 125-135.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software*, *17*(5), 1-25.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (2000). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72,* 282-307.

Rudner, L. M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice*, *20*(1), 16-19.

Schmidt, K. M., & Embretson, S. E. (2013). Item response theory and measuring abilities. In J. A. Schinka and W. F. Velicer (Eds.), Research Methods in Psychology (2nd ed.). Volume 2 of Handbook of Psychology (I. B. Weiner, Editor-in-Chief).

Schultz-Larsen, K., Kreiner, S., & Lomholt, R. K. (2007a). Mini-Mental Status Examination: A short-form of MMSE was as accurate as the original MMSE in predicting dementia. *Journal of Clinical Epidemiology, 60,* 260-267.

Schultz-Larsen, K., Kreiner, S., & Lomholt, R. K. (2007b). Mini-Mental Status Examination: Mixed Rasch model item analysis derived two different cognitive dimensions of the MMSE. *Journal of Clinical Epidemiology, 60,* 268-279.

Shallice, T., & Warrington, E. K. (1970). Independent functioning of verbal memory stores: A neuropsychological study. *Quarterly Journal of Experimental Psychology, 22*, 261-273.

Slick, D. J. (2004). Psychometrics in neuropsychological assessment. In E. Strauss, E. M. S. Sherman, & O. Spreen (Eds.), *A Compendium of Neuropsychological Tests:*

Administration, Norms, & Commentary (pp. 1-44). New York, NY: Oxford University Press.

Slick, D.J., Hoop, G., & Strauss, E. (1995). *The Victoria Symptom Validity Test*. Odessa, FL: Psychological Assessment Resources.

Slick, D. J., Hopp, G., Strauss, E., & Spellacy, F. J. (1996). Victoria Symptom Validity Test: Efficiency for detecting feigned memory impairment and relationship to neuropsychological tests and MMPI-2 validity scales. *Journal of Clinical and Experimental Neuropsychology*, *18*(6), 911-922.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, 161-169.

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology, 1904-1920*, *5*(4), 417-426.

Spencer, R. J., Wendell, C. R., Giggey, P. P., Seliger, S. L., Katzel, L. I., & Waldstein, S. R. (2013). Judgment of Line Orientation: An examination of eight short forms. *Journal of clinical and experimental neuropsychology*, *35*(2), 160-166.

Sperling, R. A., Dickerson, B. C., Pihlajamaki, M., Vannini, P., LaViolette, P. S., … & Johnson, K. A. (2010). Functional alterations in memory networks in early Alzheimer’s Disease. *Neuromolecular Medicine, 12,* 27-43.

Squire, L. R., Stark, C. E., & Clark, R. E. (2004). The medial temporal lobe. *Annu. Rev. Neurosci.*, *27*, 279-306.

Strauss, E., Sherman, E. M. S., & Spreen, O. (Eds., 2006), *A Compendium of Neuropsychological Tests: Administration, Norms, & Commentary* (pp. 1-44). New York, NY: Oxford.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*,*240*(4857), 1285-1293.

Teng, E. L., & Chui, H. C. (1987). The Modified Mini-Mental State (3MS) examination. *J Clin Psychiatry, 48,* 314-318. In Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., … & van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology, 61,* 1018-1027.

Teng, E. L., Hasegawa, K., Homma, A., Imai, Y., Larson, E., Graves, A., et al. (1994). The Cognitive Abilities Screening Instrument (CASI): a practical test for cross-cultural epidemiological studies of dementia. *Int Psychogeriatr, 6,* 45-58. In Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., … & van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology, 61,* 1018-1027.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18,* 291-307.

Thomas, M. L., & Locke, D. E. C. (2010). Psychometric properties of the MMPI-2-RF Somatic Complaints (RC1) scale. *Psychological Assessment, 22,* 492-503.

Tinsley, H. E. A., & Dawis, R. V. (1977). Test-free person measurement with the Rasch simple logistic model. *Applied Psychological Measurement, 1,* 483-487.

Tombaugh, T. N. (1996). *Test of memory malingering: TOMM*. North Tonawanda, NY: Multi-Health Systems.

Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement*, *69*(1), 50-61.

Tranel, D., & Damasio, A. R. (2002). Neurobiological foundations of human memory. In A. D. Baddeley, M. D. Kopelman, & B. A. Wilson (Eds.), The Handbook of Memory Disorders, Second Edition (pp. 17-56). West Sussex, England: John Wiley & Sons Ltd.

Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, *54*(1), 68-80.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science, 277,* 376-340.

Wade, T. C., & Baker, T. B. (1977). Opinions and use of psychological tests: A survey of clinical psychologists. *American Psychologist, 32,* 874-882. doi: 10.1037/0003-066X.32.10.874

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5,* 125-146.

Wechsler, D. (1987). *The Wechsler Memory Scale-Revised* (manual). San Antonio, TX: The Psychological Corporation. In Makatura, T. J., Lam, C. S., Leahy, B. J., Castillo, M. T., & Kalpakjian, C. Z. (1999). Standardized memory tests and the appraisal of everyday memory. *Brain Injury, 13,* 355-367.

Wechsler, D. (2001). *Wechsler Test of Adult Reading: WTAR*. Psychological Corporation.

Wechsler, D. (2009). *Wechsler Memory scale – Fourth Edition (WMS-IV): Technical and interpretive manual.* Bloomington, MN: Pearson.

Weiss, D. J. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. J. Lubinski and R. V. Dawis (Eds.), Assessing individual differences in human behavior: New concepts, methods, and findings (pp. 49-79). Palo Alto, CA: Davies-Black Publication.

Welsh, K. A., Butters, N., Mohs, R. C., Beekly, B., Edland, S., Fillenbaum, G., & Heyman, A. (1994). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): V. A
normative study of the neuropsychological battery. *Neurology, 44*, 609–614. In Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology, 17,* 380-392.

Weng, L. J., & Cheng, C. P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*(5), 697-716.

Wilkinson, G. S., & Robertson, G. J. (2006). Wide Range Achievement Test (WRAT 4). *Psychological Assessment Resources, Lutz*.

Williams, J. M. (1991). Memory Assessment Scales. Odessa, FL: Psychological Assessment Resources. In Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology, 17,* 380-392.

Wilson, B. A. (2002). Assessment of memory disorders. In A. D. Baddeley, M. D. Kopelman, & B. A. Wilson (Eds.), The Handbook of Memory Disorders, Second Edition (pp. 617-636). West Sussex, England: John Wiley & Sons Ltd.

Winegarden, B., Yates, B., Moses, J., Benton, A. L., & Faustman, W. (1998). Development of an optimally reliable short form for judgment of line orientation. *The Clinical Neuropsychologist, 12,* 311-314.

Woods, S. P., Delis, D. C., Scott, J. C., Kramer, J. H., & Holdnack, J. A. (2006). The California Verbal Learning Test – second edition: Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Archives of Clinical Neuropsychology, 21,* 413-420.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance, & R. J. Vandenberg. (Eds.). Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences (pp. 37-59). New York, NY: Routledge/Taylor & Francis Group.

APPENDIX A

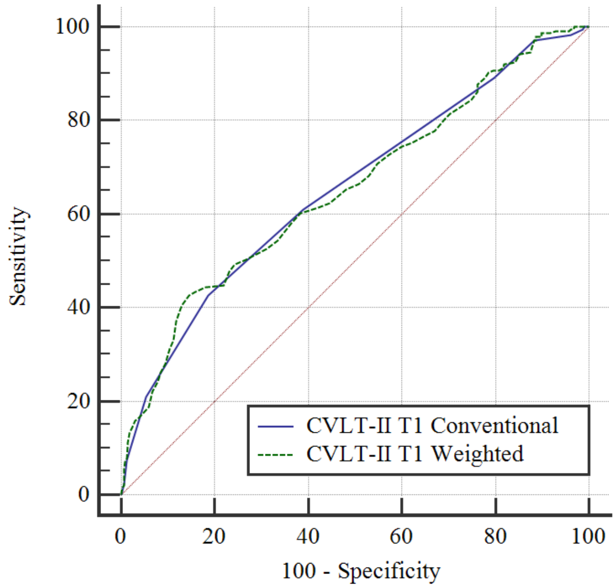


Figure 48 CVLT-II Trial 1 ROC curves for the conventional and weighted scoring approaches.

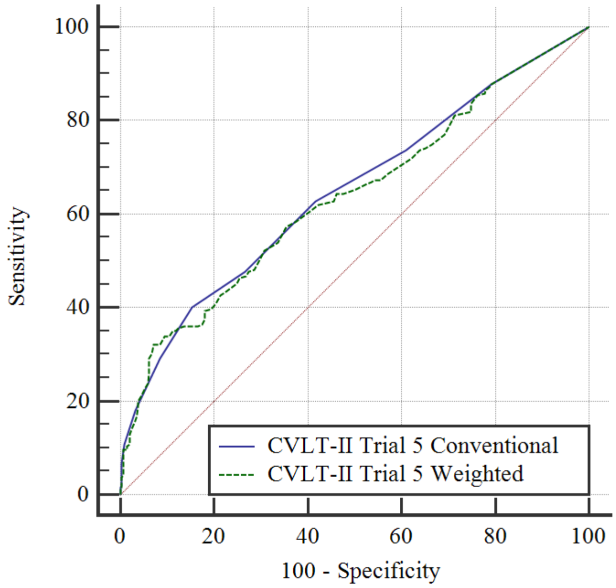


Figure 49 CVLT-II Trial 5 ROC curves for the conventional and weighted scoring approaches.

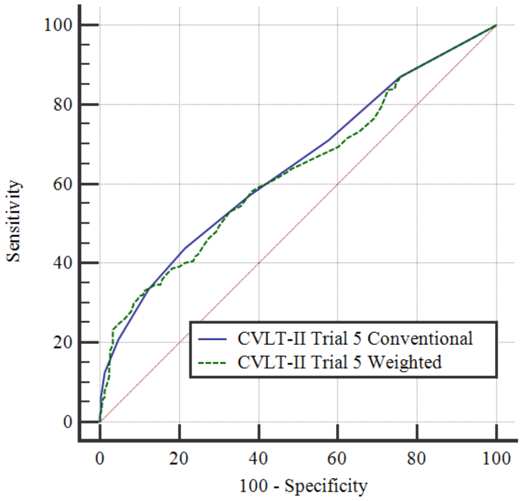


Figure 50 CVLT-II Trial 5 (14-items) ROC curves for the conventional and weighted scoring approaches.

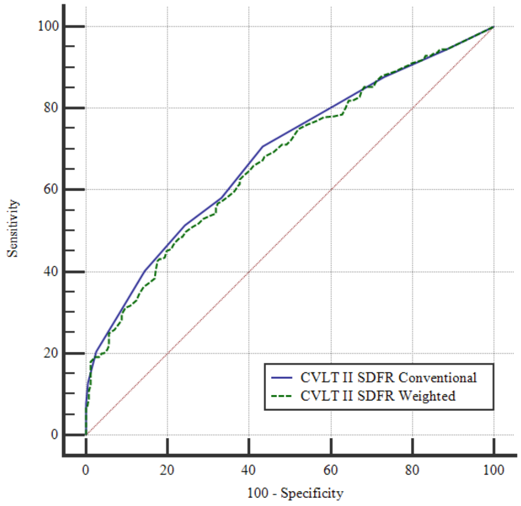


Figure 51 CVLT-II SDFR ROC curves for the conventional and weighted scoring approaches.

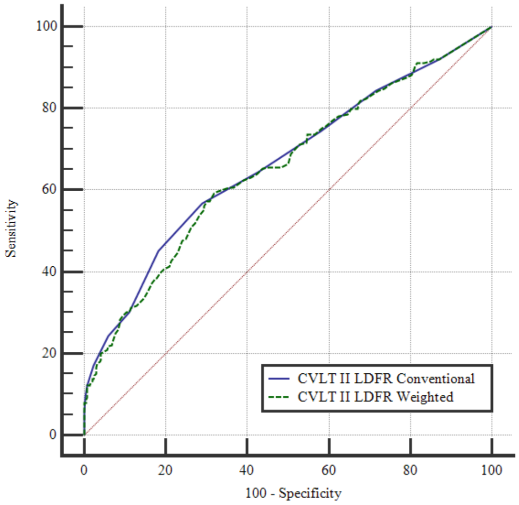


Figure 52 CVLT-II LDFR ROC curves for the conventional and weighted scoring approaches.

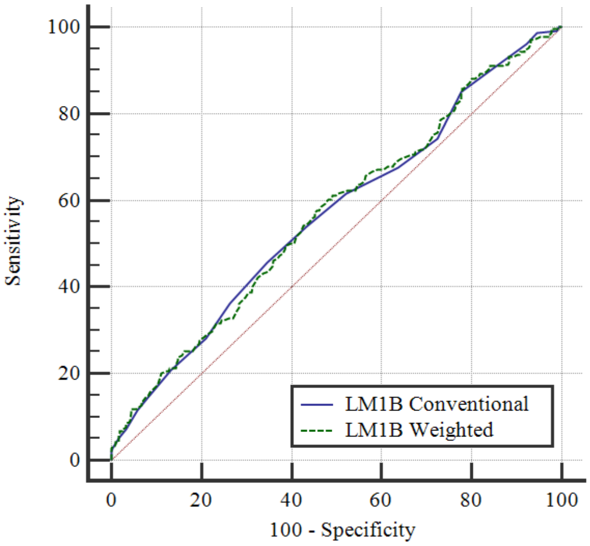


Figure 53 LM I Story B curves for the conventional and weighted scoring approaches.

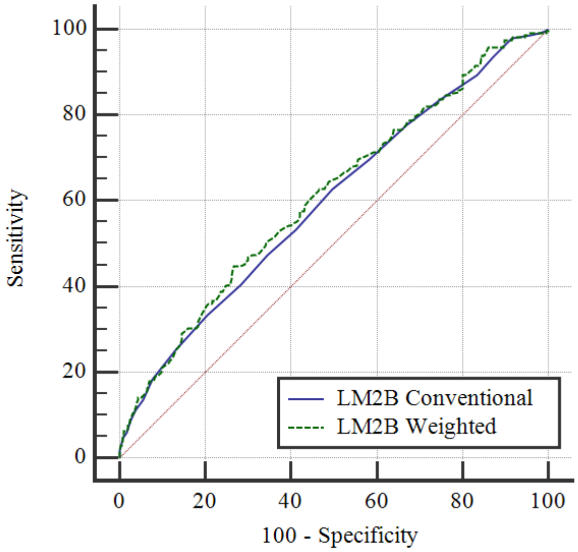


Figure 54 LM II Story B curves for the conventional and weighted scoring approaches.

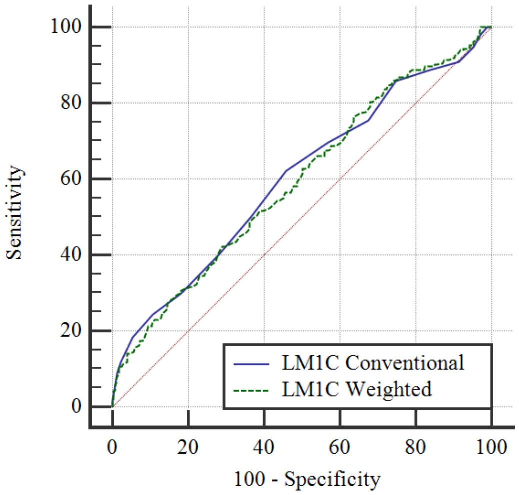


Figure 55 LM I Story C curves for the conventional and weighted scoring approaches.

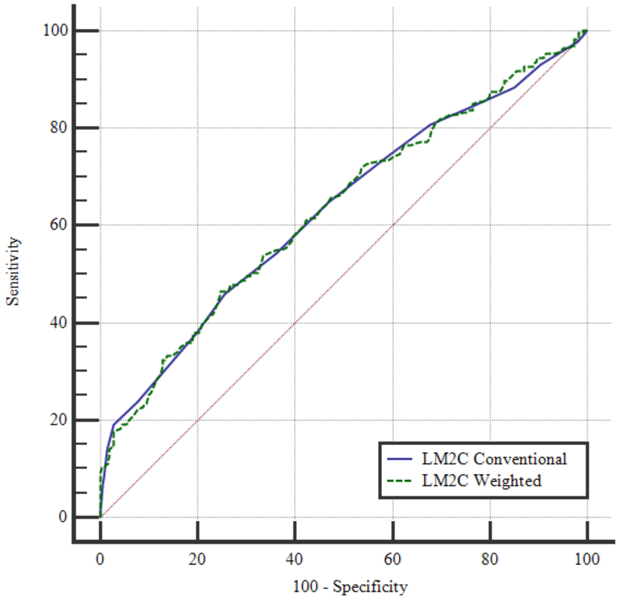


Figure 56 LM II Story C curves for the conventional and weighted scoring approaches.