SIMILARITIES AND DIFFERENCES AMONG COMMONLY USED VERBAL LIST LEARNING TESTS

_____

A Dissertation

Presented to

The Faculty of the Department

of Psychology

University of Houston

_____

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

_____

By

Chad Parker Johnson, M.A.

August, 2012

SIMILARITIES AND DIFFERENCES AMONG COMMONLY USED VERBAL LIST LEARNING TESTS

_____

An Abstract of a Dissertation

Presented to

The Faculty of the Department

of Psychology

University of Houston

_____

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

_____

By

Chad Parker Johnson, M.A.

August, 2012

Abstract

The current study identifies intra-individual performance differences on three commonly used verbal list learning tests, and it discusses implications of these differences for both clinical and research applications. The measures of interest are the California Verbal List Learning Test – Second Edition, the Hopkins Verbal List Learning Test – Revised, and The Rey Auditory Verbal List Learning Test. While each measure is classified as a verbal list learning test, differences in test structure and administration may result in variable performance within individuals. This variability has potential implications for clinical test selection under various circumstances and for utilization of the tests in research. To address questions about the similarity of these measures and comparability of scores, the author obtained scores on all three tests from a sample of 92 normal college students. In addition, learning curve characteristics, serial position effects, and semantic clustering effects were compared and contrasted across measures. Correlations for similar measures within tasks were significant, but lower than would be acceptable for alternate forms use. Differences in tasks were identified in learning curve characteristics and serial position effects. Additionally, factor structures of tasks varied significantly. Discussion of results includes exploratory explanations for some sources of variance among tests. The current study reinforces the need for neuropsychologists to carefully consider their specific task selections within the testing paradigm of verbal learning, noting the population of interest, the purpose of the evaluation, and the conceptualized construct of verbal learning from which the neuropsychologist is operating.

Table of Contents

List of Tables and Figures

Similarities and Differences among Commonly Used Verbal List Learning Tasks

Assessment of verbal learning and memory is a necessary component of almost any neuropsychological evaluation. The value of verbal memory testing stems from three characteristics of verbal memory that are relevant to the clinical neuropsychologist and consequently the patient. First, verbal memory dysfunction has devastating effects on daily living skills and personal independence as it is vital for many activities that are taken for granted by typically functioning adults. Second, verbal memory is a multi-faceted umbrella construct that involves many component processes; therefore the level of deficit in daily life and the proper intervention for impairments of verbal memory require careful evaluation of the various components (Lezak, Howieson, & Loring, 2005). Finally, verbal memory is affected by a wide variety of neurological conditions and is a common complaint and referral question for neuropsychological evaluation. These characteristics of verbal memory are also reflected in research by neuropsychologists that reveals the variety of effects that neuropathology can have on verbal memory as well as the ways in which clinical evidence of memory loss can illuminate the construct of verbal memory (Squire & Schacter, 2003).

The verbal list-learning paradigm is a commonly employed measure of memory in neuropsychological practice (Rabin et al., 2005). This testing paradigm includes a collection of tests that share specific characteristics associated with the tradition of verbal learning research. The tests of this class include assessment of learning rate, encoding efficiency, recall, retention and forgetting rate. The information is valuable to the neuropsychologist as deficits on tasks of verbal learning are associated with memory problems seen in clinical populations, including traumatic brain injury, frontal lesions, and dementia. Non-neurological variables, such as motivation and sleep deprivation, can also have a detrimental effect on performance (Drummond et al., 2000).

Lezak et al. (2005) recommend six distinct constructs that should be assessed in a comprehensive memory evaluation: orientation, prose recall, rote learning and memory, visuospatial memory, remote memory, and autobiographical memory. As such, there is currently no single neuropsychological tool that could bridge all areas of assessment in a comprehensive and valid manner. However, the verbal list learning paradigm is an excellent tool for assessing rote learning and memory and has been adopted by many neuropsychologists as the standard paradigm for assessment of verbal learning (Rabin et al., 2005). Further, many list learning tasks include all of the procedures needed to address storage and retrieval questions, including: immediate and delayed recall, retroactive and proactive interference, and recognition-based trials for those with recall below normal limits (Lezak, Howieson, & Loring, 2005). Factor analysis of the Rey Auditory Verbal List Learning Task suggests it is a measure of three factors of memory: encoding, storage, and retrieval, consistent with common neuropsychological interpretations of memory efficiency (Vakil & Blachstein, 1993). While this may not span the full gamut of verbal learning dimensions, it does provide an approach for identifying the initial stage of memory processing dysfunction.

An excellent review of available tests in the list learning paradigm has been produced by Mitrushina and colleagues (Mitrushina, Boone, Razani, & D'Elia, 2005), a work that reviews many of the most well-designed and available normative datasets as well as the unique characteristics of each measure and its relationship to neurocognitive abilities in brain damaged populations. In addition, studies of the relationship between earlier editions of some list learning tasks have found the properties of different tests to be similar, but not identical (Crossen & Wiens, 1994; Stallings, Boake, & Sherer, 1995). However, these

studies have not been replicated for newer editions or across more than two tests, and this limits our certainty in the concurrent validity of the most popular tests of the auditory verbal list learning test (VLT) paradigm.

While there is no consensus about what constitutes a VLT in contrast with other verbal learning measures such as the Selective Reminding Test or paired recall measures, the following guidelines are offered for defining this class of tests:

1. The VLT includes meaningful individual words as stimuli presented orally in a list;
2. Patient response is produced orally;
3. The list is repeated over multiple trials and the patient must provide recall after each trial;
4. Total recall of words is the primary measure of interest;
5. List repetition is presented identically following each recall session, regardless of performance.

These criteria clearly distinguish tests commonly called VLTs from other verbal learning tasks and differentiate them from other memory measures. The purpose in providing such a list of guidelines is to specify the testing paradigm more clearly in a field populated with numerous published testing tools. Other memory measures, such as story memory tasks, visuospatial tasks, and selective reminding tasks, share many characteristics with the VLTs but vary enough to be categorized separately. Those other measures often tap different aspects of memory which can be differentiated through double dissociation in specific neurological populations (Lezak, Howieson, & Loring, 2005).

Other components common to VLTs include interference trials, delayed memory trials, and recognition trials, though these vary among tests. Variables of interest in a VLT

generally include initial learning, change in number of words recalled over time (learning curve), retention over a delay, effect of interference, primacy and recency effects, semantic and serial clustering effects, and differences between free recall and recognition. However, unintended and sometimes seemingly arbitrary differences between tasks in the VLT paradigm may result in unanticipated changes in performance. These differences will alter the conclusions drawn from the measures, and that may have significant consequences for clinical recommendations and research conclusions. Study of the serial position curve in both animals and humans suggests that differences in list length and delay before response can alter the classic primacy and recency effects that initially drew cognitive psychologists to the paradigm of list learning (Wright, 1998). While list length varies across the VLTs currently under investigation, all the tests share a similar delay between initial and delayed memory trials. Homogeneity across this dimension facilitates analysis of other variables, including serial position, learning and, forgetting curves.

Brief History and Design of Common VLTs

Rey Auditory Verbal List Learning Task (RAVLT). The Test of Memory for Words (Test de Memoire des Mots) was developed in the late nineteenth century by Swiss neurologist Edouard Claparede. Consistent with other memory measures of the time, it was a list of commonly known words that the average adult should know. It was developed in French and used by both Claparede and later by his student, Andre Rey. Rey was primarily interested in the variations of intra-subject performance between recall and recognition trials, rather than interpretation of the learning curve or delayed recall effects (Boake, 2000; Rey, 1941). However, the task became popular because a wide variety of memory deficits could

be isolated through careful interpretation of trial learning. Over time the task eventually became the procedure known as the Rey Auditory Verbal List Learning Test (RAVLT).

The task changed markedly from its original format to its current incarnation. It was translated to English by Taylor in 1959 and several minor word changes were made to the list at that time (Mitrushina et al., 2005). However, most of the list was left unaltered, resulting in at least one unintended change to the nature of the test. Claparede had selected words with similar syllable structure (2 syllables per word) to be consistent with recommendations in French test development compendia of the time (Boake, 2000). The English translation includes words with both one and two syllables as a result of the direct translation.

The RAVLT has been used extensively as a measure of verbal learning and memory with a variety of minor modifications and various administration protocols (Mitrushina et al., 2005). Various writers made changes that included reversing the order in which recognition and recall trials were presented, altering the length of the delay between administration and recall, and changing the format of the recognition test. Consequently, when studying and selecting normative data from a compendium such as the one edited by Mitrushina and colleagues (2005), special care must be paid to format differences so as to determine comparability to the current administration. Due to these variations, it is difficult to generalize findings from different studies and populations. Furthermore, the common approach of generating meta-analytic normative datasets for orphaned tests, as has been done by Mitrushina and colleagues (2005), as well as by Schmidt (1996), may not provide valid regression coefficients for weighting demographic characteristics due to an overall dearth of appropriate and standardized normative datasets. However, despite its modifications and various forms, the RAVLT can be considered the prototype for later tests of verbal list

learning in neuropsychology. Additionally, the RAVLT in one form or another continues to be one of the most common memory assessment tools utilized by neuropsychologists in applied settings (Rabin, Barr, & Burton, 2005).

California Verbal List Learning Task – Second Edition (CVLT-II). The California Verbal Learning Test (CVLT) was developed by Dean Delis and colleagues as a more involved, complex, and rigorously standardized verbal list learning task (Delis, Kramer, Kaplan, & Ober et al., 2000). While several problems plagued the original test – including poor control for item selection, inflated normative sample, and the presence of a superordinate category (shopping items) – the revised edition, CVLT-II, has become one of the most popular measures of memory used by neuropsychologists in both research and practice (Rabin et al., 2005). Part of this popularity can be attributed to the CVLT-II being designed as a comprehensive memory measure that covers a variety of memory subdomains associated with neurological conditions.

The CVLT-II is the longest of the three tests being reviewed. It includes immediate recall trials, semantically cued trials, interference, delayed free and semantic recall, and a recognition trial. In addition, there is a forced-choice trial that is primarily used in studies of symptom validity (Delis et al., 2000). The CVLT-II produces 66 normed variables on the Expanded Report as well more than 260 non-normed process variables (Sherman, Strauss, & Spreen, 2006). The authors argue that these additional variables provide a basis for integrating cognitive neuroscience theory into neuropsychological assessment by allowing for better study of the process of memory deficits, though the sheer number of variables makes it difficult to formulate performance profiles (Delis, Novack, Trenerry, & Craig, 1988).

The CVLT-II as a tool is a double-edged sword due to its length and complexity. In favor of using the task is the presence of a wealth of well-normed data that allow the investigator to address immediate rote learning, five-trial learning, learning curve across five trials, proactive interference, retroactive interference, immediate free recall, immediate cued recall, delayed free recall, delayed immediate recall, and characteristics of recognition. This allows for differentiation of profiles within very specific patterns of memory deficits (Delis et al., 1991). However, the complexity of the CVLT-II also introduces two major limitations. First, early trials may "pollute" later trials by providing structure that improves subsequent retrieval. For example, Crosson et al. (1988) reported that an immediate cued recall trial improves delayed free recall to a level that exceeds immediate free recall in both normal and head-injured individuals. Second, the inclusion of semantic categories and cued retrieval provides an opportunity to use strategies that increase contextual efficiency (i.e., semantic clustering). This altered efficiency may be incompatible with the purposes of the researcher or clinician.

Hopkins Verbal List Learning Task – Revised (HVLT-R). The HVLT-R was initially designed as a shortened verbal list learning task designed to replace longer administrations of the verbal list learning paradigm, such as the CVLT-II, in dementing populations (Mitrushina et al., 2005). This change was justified by evidence of substantial fatigue effects in some populations (Brandt & Benedict, 1998). Additionally, due to floor effects on the RAVLT in the older age ranges, the authors developed the HVLT to better track the progression of dementia. The authors also noted the need for more rapid evaluation of verbal memory due to Medicare policies (Benedict et al., 1998). However, constrained variation in HVLT-R scores and modest correlation with CVLT-II scores in an Alzheimer's

sample suggests that there may be a loss of sensitivity associated with the use of HVLT-R even in dementing populations, where memory effects are likely to be significant (Lacritz, Cullum, Weiner, & Rosenberg, 2001). This includes a loss in ability to differentiate between forms of dementia and to identify specific impairments of memory (Shapiro, Benedict, Schretlen, & Brandt, 1999).

An unexpected and possibly unfortunate consequence of the shorter test has been its acceptance as a shortened verbal list learning measure for other purposes, such as pre-season concussion screening and use in acute TBI (Lovell & Collins, 2001). The pronounced ceiling effect and poor normative sample size for younger individuals hinder the utility of the test for these purposes, and its sensitivity to TBI of any severity is still unknown. The normative sample for the HVLT-R is especially small for the 16-19 (N = 29) and 20-29 (N = 84) age brackets. In addition, the HVLT-R does not assess interference effects in any way as it lacks an interference list, a common characteristic of other VLTs that allows for testing of proactive and retroactive interference effects. Despite these limitations, the HVLT-R is a shorter, faster test that may serve the purpose of verbal learning assessment and contrasts nicely with the lengthier CVLT-II for consideration in clinical and research studies.

Varying Test Characteristics

Many characteristics factor into a neuropsychologist's decision to employ one test over another. The current market for published neuropsychological tools is enormous as clinical psychology in the United States continues to develop as a medical service. Corporations have begun to take control of test development, providing powerful benefits such as rigorous standardization procedures and large sample sizes. However, along with this shift toward more complete psychometric studies in test development, the market for test

selection has become flooded with a multitude of options for many common test paradigms. Competing corporations have begun to develop similar tools in order to compete directly for support from neuropsychologists in private and hospital practices. The VLT paradigm is no exception to this trend. While the RAVLT is a test in the public domain, attempts have been made by two companies to profit from the test by producing more convenient forms of the test for sale to clinicians. In addition, one company has published a manual containing normative data and test administration procedures for the RAVLT (Schmidt, 1996). Because the CVLT-II and HVLT-R are each produced by competing companies, the clinical neuropsychologist may easily find himself/herself trying to determine which tool would be preferable in his/her practice. In addition to the three tests described herein, there are several verbal list learning tasks associated with popular memory batteries, such as the Wide Range Assessment of Memory and Learning and the Repeatable Battery for the Assessment of Neuropsychological Status, either of which may be selected to ensure homogeneity of samples between memory tasks included in the battery. If all verbal learning tasks shared similar psychometric properties, then research with any single tool could be applied to assessment with any other tool. However, evaluations of previous editions of the VLTs, including the CVLT and HVLT, suggest that these tasks produce divergent scores on corresponding measures, limiting the generalizability of findings from one test to another (Lacritz & Cullum, 1998; Crossen & Wiens, 1994). The following is a summary of key differences between the tasks that may affect both the intra-individual variation in performance and the likelihood of selection by a neuropsychologist in various clinical or research roles. These variables are also summarized in tabular form in appendix 1.

Number of Learning Trials. The number of learning trials varies between the 5 trials of the RAVLT and CVLT-II and the 3 trials of the HVLT-R. For the HVLT-R, this limits opportunities to learn material and reduces the span over which measures of learning span can occur. The learning curve itself on VLTs is generally not linear despite the linear calculations presented in the CVLT-II manual and is better expressed as a quadratic function with steep initial gains followed by a leveling or plateauing effect (van der Elst, van Boxtel, van Breukelen, & Jolles, 2005; Delis et al., 2000). Data from 1855 normal participants demonstrate that there is a plateauing effect following the third trial of the RAVLT (van der Elst et al., 2005). Due to this non-linear increase, which is a function of list length, word familiarity, interstimulus interval, and semantic similarity, it is difficult to estimate a theoretical learning curve for the missing fourth and fifth trials of the HVLT-R. However, van der Elst and colleagues have suggested that calculating a modified delta variable as a measure of the difference in recall between trials 1 and 3 on the RAVLT may serve as a better measure of rote learning as it corrects for trial 1 while minimizing ceiling effects (2005). Therefore, delta will be compared with traditional learning curve calculations of the difference between trials 1 and 5 on the RAVLT and CVLT-II as well as with the standard learning curve on the HVLT-R which has only 3 trials. These comparisons will provide information about the risk of losing information by reducing the number of trials. The primary concern is the consequence of missing informative profile effects, which may differentiate individuals as well as neurological populations.

Number of Items. All three tests contain different numbers of items. In general, previous studies have found that there is about a 1 point improvement in raw scores per trial for the CVLT-II relative to the RAVLT. This difference is consistent with the presence of an

additional item on the CVLT-II (Crossen & Wiens, 1994; Stallings et al., 1995). However, it is unlikely that the increase is a simple one-to-one relationship between score and number of items. This can be illustrated through a simple thought experiment. For the most extreme short-length lists, the addition of one item results in an average gain of one item (i.e., memorization of a single item is likely to have a mean close to one while a list of two is likely to have a mean close to two). For extremely long lists, it is unlikely that any meaningful gain will be seen between a list of 100 words versus 101. At the least, the list-length effect is likely to be parabolic or quadratic with tapering, though the presence of competing primacy and recency effects may result in a more complex pattern. Consequently, the much shorter list length of the HVLT-R, when compared to the other tasks, may not demonstrate such a clear raw score effect when applied to normal adults, whose performance falls at the high end of the range of possible scores. The impact of the shorter list on the performance of impaired individuals is still relatively unknown (Brandt & Benedict, 1998).

Familiarity of Items. The RAVLT was initially developed in French and only three items were altered in translation to common English words. In addition, the word list for the RAVLT pre-dates assessments of word familiarity and probability which have been used in developing modern VLTs (Thorndike-Long, 1944). By contrast, both the CVLT-II and HVLT-R have been developed with common categories of words in mind and analyses of word familiary/commonality were used to identify words with similar rates of usage and to eliminate likely prototype category members so that true memory of items could be differentiated from commonly confabulated items (Benedict et al., 1998; Delis et al., 2000). The consequences of these differences between the RAVLT and newer tests can most clearly be seen in semantic clustering, which is strongly primed by both the presence of categories in

both the HVLT-R and the CVLT-II. Word associations also may increase the use of cued-recall in the CVLT-II, and on recognition tasks, they may increase the likelihood of semantically related intrusions. Conversely, the RAVLT features weakly associated target words, reducing the likelihood of semantic intrusions.

Semantic Relationships. The presence of semantic relationships increases the efficiency of encoding in intact adults (Lezak, Howieson, & Loring, 2005). However, the presence of similar items in the list also increases the possibility of intra-list interference effects as described earlier, which may result in a greater rate of semantically related intrusions in recall and paradoxical forgetting for delayed recall in clinical populations, such as brain injured patients, who benefit from provided structure (Crosson, Novack, Trenerry, & Craig, 1988; Runquist, 1966). However, semantic clustering effects provide an additional means of measuring the executive functions underlying memory efficiency, and each of the three tests has dealt with this component differently. The RAVLT includes a few incidentally related items which do not lend themselves to formalized analysis. The HVLT-R includes semantic categories but never structures the clustering for the participant by cuing or naming the categories. The CVLT-II includes both immediate and delayed semantic cued recall, which allows for structuring of the items into semantic categories at the cost of tampering with the participant's natural encoding strategy, alleviating working memory limitations, and increasing the use of semantic strategies. A study of the previous version of the CVLT demonstrated frequent use of semantic clustering strategies, though without an increase in overall recall in healthy participants (Shear, Wells, & Brock, 2000). However, there is a possibility that the effect of semantic clusters, and the manner in which they are

handled in the newer edition, may have an effect on the total words learned by the participant.

Interstimulus Interval. The interstimulus interval is the period of time between stimulus offset and onset of the next stimulus. The RAVLT instructions dictate a one second interstimulus interval. The CVLT-II instructions state that words should be read at a rate of "about one per second", dictating an approximate time in terms of stimulus onset to stimulus onset instead of a true interstimulus interval. The result is a recommendation for a similar interstimulus interval, though with less emphasis on the rigor of timing. By contrast, the HVLT-R instructions indicate that the interstimulus interval should be two seconds, allowing greater time for possible rehearsal, decay, or manipulation. The time between stimulus presentations has been shown to interact with age and affect the number of words recalled. Longer interstimulus intervals benefit older populations, but appear to not affect performance for younger adults (Meijer et al., 2006). However, this effect has only been reported for research measures with varied interstimulus intervals and has not been demonstrated in existing published VLTs.

Interference Trial. Both the RAVLT and CVLT-II include an interference trial that allows for assessment of both proactive and retroactive interference on immediate memory. The HVLT-R lacks this feature and, as such, falls short of the required procedures recommended by Lezak et al. (2005). However, as described above, having an interference trial introduces complications for the other two tasks insofar as it contaminates the measurement of decay and recognition on later trials. Despite the possible benefits of omitting an interference trial, the interference trial allows for measure of proactive and retroactive interference, and the presence of semantically related and unrelated foils on the

interference list allows for process level assessment of memory integrity, allowing the neuropsychologist to classify false positives as related terms or random guesses based on foil similarity (Delis et al., 2000).

Cost. While unrelated to the basic psychometric properties of the tests, cost has become an increasingly powerful factor in the determination of test use by both researchers and clinicians, and this may affect the frequency with which a particular measure is used with certain populations. The RAVLT is in the public domain and is appealing to clinicians and researchers who wish to minimize cost. In contrast, the CVLT-II and HVLT-R kits cost several hundred dollars each with additional cost for forms. While the impact of cost is difficult to assess, it provides another non-random characteristic of test selection that may determine who receives which test.

Normative Samples. The RAVLT has a large collection of samples to select from when choosing a normative comparison group and these have been collected into several compendia that describe specific characteristics of the samples (Mitrushina et al., 2005; Schmidt, 1996). However, as these norms were primarily drawn from local convenience samples, or collected incidentally as part of a larger study, the size and representativeness of any RAVLT sample is likely to be limited. Conversely, the CVLT-II and HVLT-R have corporate-funded normative groups that were selected specifically for that purpose and generally have better sampling characteristics. The effects of this discrepancy have been illustrated by different patterns of results for the earlier version of the CVLT and the RAVLT in both normal and TBI groups whereby similar raw score performance resulted in different categorization of memory impairment (Crossen & Wiens, 1994; Stallings et al., 1995).

Study Justification

The current study addresses some of the variations described above. Specifically, it serves three primary goals for the application of VLTs by the clinical neuropsychologist: establishment of norms within each test for the young adult age range based on standardized administration of the tests; replication and expansion of studies of previous editions of the tests that allow comparison of test characteristics within subjects; and investigation of differences between tests.

Normative dataset development. With the exception of the rigorous norms developed for the CVLT-II, the tests have not been widely studied in the normal, young adult population. Despite the frequent use of college students as convenience samples in a number of studies, no single study has established a well-standardized dataset for this population. In particular, RAVLT studies rarely include this age range, and existing meta-analytic datasets do not have appropriate power to be used in establishing valid norms for young adults. Accordingly, the norms generated by the two most notable sources are discrepant from each other (Mitrushina et al., 2005; Schmidt, 1996). The HVLT-R similarly lacks a large normative dataset in the young adult age range, owing largely to the selection of the sample to serve older populations, consistent with its original developmental purpose.

Comparison of test characteristics. The current study also expands on previous work designed to determine if intra-individual test performance differs among VLTs with various test characteristics (Crossen & Wiens, 1994; Stallings et al., 1995). The current approach is unique in that it approaches these tests from a pragmatic perspective. That is, the current study focuses on specific individual variations in performance that arise from differences in test characteristics rather than differences that arise from experimentally

varying characteristics within these established tests. Similarities and differences in performance on these three popular VLTs will inform clinicians and researchers alike on the feasibility of considering the three VLTs as comparable tasks. The comparison of scores will also provide a needed measure of concurrent validity and reliability, as called for by Sherman, Strauss, and Spreen (2006).

Explanation of variance. Finally, the current study attempts to explain the variance in test performance on the basis of existing test characteristics. This process will provide initial indications of sources of variance, such as semantic clustering and list length, on overall performance in normal controls. This will inform test selection when it is necessary to assess neurological populations known to have specific weaknesses that may be reflected in measures such as serial position and forgetting curves. While there are advantages to this approach, there are also limitations. Additional investigation of these effects will require the inclusion of alternative measures of memory and executive functioning.

The specific questions addressed by this study have been broken into three levels of organization that are broadly consistent with the purposes listed above: Preliminary analyses, primary analyses, and supplemental analyses. Preliminary analyses include methodological checks regarding order effects, demographic information about the groups, and relevant normative data tables to assist in future analyses with each of the tasks. Primary analyses include comparisons of commonly used scores for the three tests and differences in the performance of individuals across the tests. Supplemental analyses are intended to carry out exploratory level analyses of likely sources responsible for identified differences and provide a factor analysis of the three tests to identify different constructs tapped by the VLT paradigm.

The specific objectives of each level of analysis necessary to achieve the study goals are outlined below.

Preliminary Analyses

1. In order to collapse the three groups into a large normative dataset, it is necessary to test for order effects in primary scoring variables (e.g., trial 1 total, last trial total, total learning, learning curve, delayed recall). Therefore, initial analyses targeted order effects among the three orders of administration.
2. Secondary order-effects analysis considered if there were order effects for less commonly used scoring variables (e.g., intrusions, clustering strategies, recognition). While these would not preclude the use of the collapsed dataset as a normative base for scoring purposes, they would suggest that caution must be exercised as expanded data may be corrupted by order effects.
3. Resulting means and standard deviations for each measure, grouped by gender, were included along with demographic information in each cell, to provide additional normative data for future study and clinical use.

Primary Objectives

1. Comparable scores from each of the three tests were compared via Pearson correlations. It was expected that most variables would correlate significantly, consistent with the similar format of the three tests, but that correlations would fall too low to consider the tests as alternate forms of one another and raise questions as to sources of variation between the tests.

2. Trend analysis was employed to explore similarities and differences in the pattern of performance across learning trials.
3. The serial position effect was calculated for each of the three measures to see if task characteristics had an effect on this generally robust feature of list learning.

Supplemental Objectives

1. Serial clustering effects for the three measures were evaluated with the expectation that greater serial clustering would be seen overall on the RAVLT, due to the lack of cuing or overt categories for list items.

2. Similarly, the semantic clustering variable on the CVLT-II was expected to explain differences in serial clustering strategies between the CVLT-II and RAVLT, indicating that one strategy is generally independent from another.

3. The use of semantic clustering on the CVLT-II was expected to explain differences in overall performance between the CVLT-II and RAVLT, as those who rely on serial clustering strategies were expected to perform similarly on both tasks, while those who prefer semantic clustering were expected to perform better on the CVLT-II than on the RAVLT.

4. With the addition of more composite variables derived from existing scores, factor analysis of the variables from the three tasks was expected to yield a factor structure consisting of four factors (Attention Span, Learning Efficiency, Delayed Memory, and Inaccurate Memory). This factor structure was obtained in an earlier study of CVLT-II in normal controls (Donders, 2008).

Method

Participants

Participants were recruited from the University of Houston following guidelines for recruitment and study under the University of Houston's Institutional Review Board, the Center for the Protection of Human Subjects. All undergraduate participants were offered extra credit in psychology courses for participation in the study. Inclusion criteria included all typical adult (18 years of age or older) undergraduates. Exclusion criteria included age less than 18 years, severe to profound sensory loss, lack of English fluency, and history of significant psychiatric or neurological disorder. Participants were recruited via volunteer flyers, classroom announcements, and sign-up sheets within the Department of Psychology. 102 participants were recruited into the study through this method. Two withdrew prior to consent procedures when the study was described due to failure to be able to commit to required follow-up sessions. Of the remaining 100 participants, seven failed to complete the required follow-up sessions and one was excluded due to irreconcilable administration error by an examiner. Additionally, three participants did not complete the final recognition subtest of one measure (RAVLT), but were included in the final sample for all other analyses. The final sample therefore includes 92 participants with generally even division across the three conditions, though only 89 participants for recognition trial analyses. Due to slight variance in participant attrition by study cell, the final breakdown of participants included 29, 31, and 32 participants in the three cells. The flowchart of inclusion/exclusion and cell assignment is illustrated in appendix 2.

Demographics. Participant demographics captured included sex, age, years of education, and NIH classification of race/ethnicity as these variables are recommended

sample characteristics by Mitrushina et al. (2005) for normative datasets. The overall mean age of participants was 22.48 years with a standard deviation of 6.34 years. However, the distribution is positively skewed with 11 participants older than 30 years and young participants bound at 18 years. While most level of education values ranged from 12 to 16 years of previous education, 5% of participants had greater than 16 years of education and one outlying post-doctoral student presented with 22 years of prior education. Overall, years of education values were restricted due to the college student sample. Female participants greatly outnumbered male participants with 71% of the sample being female. These demographics are summarized along with cell level analyses of differences in appendix 3.

Procedures

Test Examiners. Examiners included a team of eight undergraduate research assistants, a designated undergraduate research coordinator, and the principal investigator [CPJ]. Research assistants were provided with several hours of training across days on each test and at least three practice administrations with each tool to determine proficiency before seeing any participants. Due to the value of measuring test characteristics between tasks, consistent administration practices were held paramount in data capture. Therefore, administration checks were performed by the principal investigator and research coordinator to guard against examiner drift. Additionally, when available, the research coordinator checked over administration procedures immediately following administration to identify any validity threatening violations in administration.

Study Design. Due to concerns regarding order effects and the need to analyze the possible practice effects associated with position, a Latin Squares repeated measures design was implemented. While previous studies have not identified practice effects associated with

earlier editions of two of the VLTs under examination (Crossen & Wiens, 1994; Stallings et al., 1995), due to the increase in number of tests, changes associated with updates, and possible carryover test effects (ex., semantic clustering strategies or increased intrusions), it is worth investigating the possibility of order effects and protecting against possible meta-cognitive practice effects for later trials.

Participants were required to attend three separate testing sessions with a single VLT administered at each session as dictated by group membership in the study design (Appendix 4). Participants were tested over the course of three test sessions held between 3 and 9 days from each other, resulting in a possible range of 6 to 18 days for total administration. During each session, a single VLT was administered, whose order was selected randomly (Urbaniak & Plous, 2011; see Appendix 5 for randomization scheme). During the initial session, participants were informed of the purpose and nature of the study as part of the consenting procedures. During follow-up sessions, a brief reminder of the nature of the study was presented to allow for several minutes to allow the participants to re-acclimate to the testing environment before beginning the task. Each session took between 35 and 55 minutes to complete including the 20 minute delay task and the final session included an additional 15 minutes to complete a computerized symptom validity task to address concerns regarding motivation to perform in participants. A sample flowsheet of time commitment per participant is included in appendix 6.

The University of Victoria Symptom Validity Task (Slick, 1994), was administered to verify sufficient motivation to perform by safeguarding against sub-optimal effort. Participants who failed to meet a required criterion of 18 correct hard items, placing their performance firmly in the invalid range for effort, were excluded from the final analysis,

consistent with work completed by previous authors (Loring, Lee, & Meador, 2005). However, no participants in the final sample failed to meet the sufficient motivation cutoff, though five participants fell in the questionable validity range. Additionally, due to technical difficulties in administration of the Victoria Symptom Validity Test, five participants lack complete motivation testing data, but were still included in the final analysis.

General Administration Guidelines. All tasks were administered as specified in their respective manuals. In the case of the RAVLT, the administration guidelines proposed by Schmidt were used as these provide a published, standardized set of instructions associated with a normative dataset for comparison. Additionally, the administration instructions in this handbook are identical to those provided in other common sources for the task (Lezak, 2004; Strauss, Sherman, & Spreen, 2006). All required tasks were included for all tests, but optional subtests, such as the CVLT's forced choice recognition, were not administered. To avoid the possibility of verbal interference during the 20 minute delay on each task, participants were presented with a series of visuospatial Sudoku puzzles to complete during the 20 minute period. At the end of 20 minutes, participants were presented the delayed recall portion of testing, regardless of how many puzzles they had completed. Sudoku puzzles were described as measures of cognitive processing speed to promote participant effort.

Scoring Guidelines. Raw scores were calculated for several commonly used composite variables across measures. Due to some test variation, additional composites were also generated for task comparison purposes. For example, the HVLT-R has a three trial learning curve; therefore a three trial learning curve variable was also generated for the RAVLT and CVLT-II. Examiners scored the measures and all protocols were double-scored

by the principal investigator and the research coordinator to ensure fidelity in scoring. All data from the two sources was transposed to an Excel spreadsheet and discrepancies between scores were identified between the scorers. Scores with inter-rater discrepancies were re-evaluated in raw form by the principal investigator to ensure accurate score calculation. Inter-rater concordance fell at 96% between the principal investigator and the research coordinator and this was deemed within an acceptable range to ensure that multiple raters resulted in accurate scoring following rescoring of discrepant values. Target variables are defined below and summarized in tabular form in appendix 7.

Variable calculations.

Initial Learning Trial Variables.

Trial 1 Correct, Trial 2 Correct, Trial 3 Correct, Trial 4 Correct, Trial 5 Correct. Due to the format of the tests, all three measures include trial 1 through 3 correct measures which include the raw number of correct words recalled. The CVLT-II and RAVLT also have Trial 4 and 5 correct variables with similar characteristics. The total learning variable is the sum of all trial correct variables for the measure. For all three measures, there is normative data to convert scores on individual trial and total learning variables.

Trial 1 through Trial 5 Repetitions. This variable represents the raw number of times that a target word is repeated by the participant during an administration as though it had not already been spoken. Total repetitions is the sum of all trial level repetitions. Only the CVLT-II provides normative data for this variable.

Trial 1 through Trial 5 Intrusions. This variable represents the raw number of times that a word not present on the target list is reported as though it were a member item. Total

intrusions is the sum of all trial level intrusions. Only the CVLT-II provides normative data for this variable.

Trial 1 through Trial 5 Semantic Intrusions. This variable represents the number of items that are intrusions as defined above which also fit into one of the semantic categories on the CVLT-II or HVLT-R. This variable cannot be calculated for the RAVLT due to the lack of semantic categories. Total semantic intrusions is the sum of all trial level semantic intrusions. No measure provides normative data for this variable.

Trial 1 through Trial 5 Serial Clusters. This variable represents the raw number of target words reported which were immediately followed by a target word that either preceded or succeeded it in the original presentation list. This variable can be calculated for all three measures. Total serial clusters is the sum of all trial level serial clusters. Only the CVLT-II provides normative data for this variable.

Trial 1 through Trial 5 Semantic Clusters. This variable represents the raw number of target words reported which were immediately followed by a target word that also fell into the same semantic category. This variable can be calculated for the CVLT-II and HVLT-R. Total semantic clusters is the sum of all trial level semantic clusters. Only the CVLT-II provides normative data for this variable. However, the CVLT-II version of this measure is based on a different formula which is not comparable to the current method.

Serial Position Effects.

Primacy Effect. This variable represents the total raw number of times that the first four target words as ordered on the presentation list on any of the three lists were reported during a learning trial of that task. Therefore, this variable can range from 0 to 20 on the

CVLT-II and RAVLT (4 target words x 5 trials = 20) and 0 to 12 on the HVLT-R (4 target words x 3 trials = 12). Only the CVLT-II provides normative data for this variable.

Recency Effects. This variable represents the total raw number of times that the last four target words as ordered on the presentation list on any of the three lists were reported during a learning trial of that task. Therefore, this variable can range from 0 to 20 on the CVLT-II and RAVLT (4 target words x 5 trials = 20) and 0 to 12 on the HVLT-R (4 target words x 3 trials = 12). Only the CVLT-II provides normative data for this variable.

Middle Effects. This variable represents the total raw number of times that any target words not in the previously defined primacy or recency regions were reported during a learning trial of a task. This variable can range from 0 to 40 on the CVLT-II (8 target words x 5 trials = 40), 0 to 35 on the RAVLT (7 target words x 5 trials = 35), and 0 to 12 on the HVLT-R (4 target words x 3 trials = 12). Only the CVLT-II provides normative data for this variable.

Interference Trial.

Interference Trial Correct, Repetitions, and Intrusions. These variables are calculated the same as for individual learning trials, but with regard to the target words on the interference list. These variables can only be calculated for the CVLT-II and RAVLT as the HVLT-R does not have an interference trial. Only the CVLT-II provides normative data for this variable.

Interference Trial A-List Intrusions. This variable represents the total number of target words from the learning trials that are reported during the interference trial. This variable can only be calculated for the CVLT-II and RAVLT as the HVLT-R does not have an interference trial. Only the CVLT-II provides normative data for this variable.

Immediate Recall.

Immediate Recall Correct, Repetitions, and Intrusions. These variables are calculated the same as for individual learning trials. This variable can be calculated on the CVLT-II and RAVLT. Both the CVLT-II and RAVLT provide normative data for this variable.

Trial B-List Intrusions. This variable represents the total number of target words from the interference trial that are reported during the immediate recall trial. This variable can only be calculated for the CVLT-II and RAVLT as the HVLT-R does not have an interference trial. Only the CVLT-II provides normative data for this variable.

Delayed Recall.

Delayed Recall Correct, Repetitions, and Intrusions. These variables are calculated the same as for individual learning trials. This variable can be calculated on all three measures. All three measures provide normative data for this variable.

Recognition.

Recognition Correct. This variable is the total number of correct identifications of target words when presented in the recognition trials. This variable can be calculated for all three measures. Only the CVLT-II and HVLT-R provide normative data for this variable.

Recognition False-Positives. This variable is the total number of incorrect recognitions of a non-target word as though it were a target word. This variable can be calculated for all three measures. Only the CVLT-II provides normative data for this variable.

Composites.

Learning Curve 3. This learning curve demonstrates the linear gains made between trial 1 and trial 3 correct. It is calculated as the difference between trial 1 and trial 3 correct. It can be calculated for all three measures. There is no normative data for this variable.

Learning Curve 5. This learning curve demonstrates the linear gains made between trial 1 and trial 5 correct. It is calculated as the difference between trial 1 and trial 5 correct. It can be calculated for both the CVLT-II and RAVLT. In this form, it does not have normative data. The CVLT-II provides a formula for calculating a learning curve for which there is normative data, but it does not lend itself easily to inter-test comparisons.

Semantic Clustering Ratio. This ratio is calculated as the total number of semantic clusters summed across all learning trials divided by the total number of target words identified across all learning trials. It can be calculated for the CVLT-II and HVLT-R, but does not have normative data. It is intended for inter-test comparison purposes.

Serial Clustering Ratio. This ratio is calculated as the total number of serial clusters summed across all learning trials divided by the total number of target words identified across all learning trials. It can be calculated for all three measures, but does not have normative data. It is intended for inter-test comparison purposes.

Test Reliability Characteristics. The RAVLT moderate levels of test-retest reliability when alternate forms of the task are used, falling in the .61 to .86 range after one month of testing on primary learning and retention measures (Delaney et al., 1992). The relationship of the RAVLT with the previous edition of the CVLT has already been introduced, with correlations falling in the .3 to .5 range (Crossen & Wiens, 1994). Test-retest reliability of the CVLT-II is variable, with total learning trials correlations as high as .82, but also with poorer correlations on other test variables, such as a .27 correlation for

learning curve over a 21 day delay (Delis, Kaplan, Kramer, & Ober, 2000). Reliability coefficients for the HVLT-R range from .39 for recognition scores to .74 for total recall (Brandt & Benedict, 2007). However, it should be noted that this analysis was only performed on an elderly sample. The correlation between total learning scores on the original CVLT and HVLT fell at .74 (Lacritz & Cullum, 1998). As noted previously, no studies have yet compared the more recent versions of these tasks.

Statistical Analyses. The first two preliminary of the study required order analyses of the measures. To determine if the demographics of the cells varied, an ANOVA was employed to examine differences in age or years of education between cells. Additionally, chi-square tests were employed to examine differences in frequency of remaining categorical demographic variables, including sex, race, and ethnicity between the three groups.

Following analysis of cell differences, order analyses can be performed to accomplish preliminary analyses. Order effects may arise from one of two possible sources of variation between administrations. If order effects are consistent between forms, improving uniformly across all three conditions based only on location, this is suggestive of practice effects, whereas specific cell order effects may imply meta-memoric effects (Brandt & Benedict, 1998; Crawford et al., 1989; Delis et al., 2000). To address these effects, an ANOVA of group membership was performed to assess the effect of test order on individual variables. Variables of interest included trial 1 correct, total learning, and delayed recall of the three measures. These were selected for initial analysis as these are primary scoring variables and main source of assessment in previous analyses (Crossen & Wiens, 1994; Stallings et al., 1995). Protected Fisher's Least Significant Difference comparisons were planned to address the source of order effects via linear contrast effects in the eventuality that the overall

ANOVA effect of group membership bore out. Following this analysis, a similar model was generated to investigate less commonly used scoring variables. These variables include number of intrusions, clustering strategies, and recognition scores. Following these analyses, means and standard deviations of all measures were generated by gender to provide normative data for tasks in the 18 to 29 year old range.

To address the primary objectives of the study, a large Pearson correlation matrix was initially generated from the collapsed total dataset to compare intra-individual performance on trial correct performance, total learning, learning curve, immediate recall, delayed recall, recognition correct, and recognition false-positives across measures. Following this analysis, trend analysis was performed via repeated measures MANOVA to identify differences between learning curves across the three measures for the first three learning trials. A second repeated measures MANOVA was necessary to investigate differences between the CVLT-II and RAVLT across all five learning trials due to the lack of fourth and fifth trials in the HVLT-R. To compare the serial position effect across measures, a repeated measures MANOVA was run with the serial position variables of each measure (primacy, middle, recency effects). The first level of analysis indicated whether or not serial position effects were present in the measures (i.e., if the primacy and recency scores were significantly greater than the middle score) while second level analysis indicated whether or not the effect varied across tests.

Supplemental analyses investigated the effect of strategy variables on overall performance and explored common factors shared between tests. Serial cluster ratios were compared across measures via repeated measures MANOVA. Using Pearson correlations between serial and semantic clustering ratios, the relationship between strategies on the

CVLT-II was clarified. Additionally, the CVLT-II semantic clustering ratio was included as a predictor of the difference in RAVLT and CVLT-II total learning in a general linear model to explain the relationship of alternative strategy options on participant performance variance. Finally, the effect of the CVLT-II semantic clustering ratio on differences in performance variables between the CVLT-II and RAVLT was investigated via general linear model.

Three separate exploratory factor analyses (EFA) were run to investigate factor structure across the three tests followed by a second order EFA using the defined factors from the three first-order analyses. The purpose of this analysis was to determine if the HVLT-R and RAVLT share a similar factor structure to the CVLT-II. Authors of the CVLT-II argue that it is a more complete measure of variable verbal memory functions because confirmatory factor analyses of the CVLT-II factor structure suggests 4 separate factors (Attention Span, Learning Efficiency, Delayed Memory, and Inaccurate Memory) while previous evaluations of the original HVLT and the RAVLT have generally only supported a two factor structure (Immediate Memory & Delayed Memory; Vakil & Blachstein, 1993; Donders, 2008). The current approach draws from previous analyses of factor structure presented by Vakil and colleagues' work with the RAVLT and extends this approach to include primary variables from each of the three tasks. The RAVLT analysis includes the initial trial learning (trial 1 correct), best learning trial (trial 5 correct), proactive interference (interference trial correct), immediate recall correct, delayed recall correct, the serial clustering ratio, and recognition correct variables. Initial factor structures of the CVLT-II and HVLT-R were calculated using similar variables where possible to provide a direct comparison between tasks, but expanded models were also generated and were used to

generate factor structures for those measures. The CVLT-II analysis includes the initial trial learning (trial 1 correct), best learning trial (trial 5 correct), proactive interference (interference trial correct), immediate recall correct, delayed recall correct, the serial clustering ratio, and recognition correct variables, but also includes the semantic clustering ratio and both the immediate and delayed cued recall trials. Similarly, the HVLT-R contains those variables it shares with the RAVLT (initial trial learning (trial 1 correct), best learning trial (trial 3 correct), delayed recall correct, the serial clustering ratio, and recognition correct variables), but also includes the semantic clustering ratio. Resulting factors from the initial factor analysis were submitted to a second-order analysis of construct factors shared across measures to identify construct level similarities and differences inherent to the various scores obtained through each of the VLTs.

Procedures for all factor analyses included the following steps. First, a principal factors analysis was performed. Prior communalities were estimated via the squared multiple correlation method. Factors were retained until 100% of common variance was accounted for by the sum of the initial communality values. The oblique promax rotation method was used to estimate standardized regression coefficients of factors, which involves a primary orthogonal varimax rotation followed by an oblique rotation to allow for the expectation that factors would correlate with one another. The second-order factor analysis of underlying task related factors was performed in the same manner, but the method of factor retention was set for an expected three-factor model, consistent with the most complex of the underlying component models.

Results

There were no significant differences in age, years of education, sex, race, or ethnicity between cells. Results of the relevant ANOVA and chi-square analyses are presented in appendix 3. Distributions of primary scoring variables for all three tests were generally normal, with minimal skew. However, intrusion, repetition, and false alarm scores on all three tests demonstrated highly skewed distributions and analyses of comparisons between these tasks should be interpreted cautiously.

Preliminary Analyses

1. Results of the ANOVAs for order effects and means of all three order cells are presented in appendix 8. No order effects were identified on primary scoring variables, including trial level correct responses, total learning, immediate recall, or delayed recall variables on any measure.
2. However, intrusions on immediate recall for the CVLT-II were significantly different across the three groups, $F(2, 89) = 3.97$, $p = 0.02$. Second administration of the CVLT-II resulted in significantly fewer intrusions on immediate recall than did first or third administration, $F(1, 89) = 7.24$, $p = 0.01$. No other variables demonstrated vulnerability to order effects.
3. Normative data on the three VLTs for males and females in the 18-29 year old range are provided in appendices 9-11.

Primary Objectives

1. Table 1 shows correlations for trial correct performance, total learning, learning curve, immediate recall, delayed recall, recognition correct, and recognition false-positives.
2. A 3 x 3 repeated measures MANOVA for learning curves of the three measures over the first three trials revealed a significant interaction of test with trial, $F(4, 364) = 9.11$, $p < 0.0001$. Further investigation of this effect revealed that the HVLT-R learning curve differed from the CVLT-II between the second and third trial, $F(1, 91) = 18.55$, $p < 0.0001$. In addition, the HVLT-R learning curve differed from the RAVLT learning curve between the first and second trial, $F(1, 91) = 6.00$, $p = 0.016$, as well as between the second and third trial, $F(1, 91) = 7.51$, $p = 0.0007$. No differences were identified between the CVLT-II and RAVLT learning curves. The difference in learning curves is illustrated in figure 1.
3. A 3 x 3 repeated measures MANOVA of the serial position effect for the three VLTs revealed a significant test by position interaction, $F(4, 364) = 7.82$, $p < 0.0001$. Further investigation of this interaction revealed that the CVLT-II yielded expected primacy, $F(1, 91) = 67.85$, $p < 0.0001$, and recency effects, $F(1, 91) = 51.57$, $p < 0.0001$, and the RAVLT also yielded expected primacy, $F(1, 91) = 98.3$, $p < 0.0001$, and recency effects, $F(1, 91) = 149.71$, $p < 0.0001$. These interactions were not significantly different from one another, but the HVLT-R primacy, $F(1, 91) = 5.40$, $p = 0.022$, and recency ratios, $F(1, 91) = 28.68$, $p < 0.0001$, were significantly different from the CVLT-II and RAVLT ratios. Additionally, while the HVLT-R primacy ratio was significantly different than the middle ratio, $F(1,91) = 23.23$, $p < 0.0001$,

the HVLT-R recency ratio was not significantly different than the middle ratio. The interaction of test by effect can be visualized in the serial position effect presented in figure 2.

Supplemental Objectives

1. The RAVLT produced overall higher rates of serial clustering strategy than did either the CVLT-II, $F(1,91) = 78.22$, $p < 0.0001$, or the HVLT-R, $F(1,91) = 66.17$, $p < 0.0001$. Additionally, the CVLT-II semantic and serial clustering strategy variables were inversely correlated, $r(90) = -0.73$, $p < 0.0001$.

2. As an indication of differences in the use of serial clustering between the RAVLT and CVLT-II, the semantic clustering variable was a significant regressor, $b = .32$, $t(90) = 2.75$, $p = .007$, accounting for a small but meaningful amount of variance in serial strategy use across measures, $R^2 = .08$. This finding indicates that there is greater intra-individual difference in the use of serial clustering with more serial strategies on the RAVLT than the CVLT-II as semantic clustering increases on the CVLT-II.

3. As a predictor of overall performance differences between the two measures, the semantic clustering variable from the CVLT-II also predicted differences in overall learning between the two measures, $b = 21.56$, $t(90) = 3.96$, $p = .0001$, $R^2 = .15$. Therefore, the more semantic clustering used on the CVLT-II, the greater the difference in total learning score between the CVLT-II and RAVLT, with greater performance on the CVLT-II.

4. First round exploratory factor analysis of the structure of the three measures following guidelines presented by Vakil and Blackstein (1993) resulted in two-factor

solutions for the RAVLT and CVLT-II, and a one-factor solution for the HVLT-R. The authors of that study selected trial 1, last trial, immediate recall, delayed recall, interference trial, recognition, and a strategy variable for serial clustering in their model. Due to the lack of semantic categories on the RAVLT, the Donders model could not be fit for this task (2008). To create comparability, the same method was applied to the VLTs in the study. This factor analysis excluded variables from the CVLT-II and HVLT-R that are not present on the RAVLT. This initial analysis indicated that factor structures between the CVLT-II and RAVLT were similar when only variables associated with both tests were included. When additional variables from the HVLT-R and CVLT-II were added to the model, the analysis produced a two-factor solution for the HVLT-R and a three-factor solution for the CVLT-II. The factors were labeled following recommendations by Vakil and Blackstein in their labeling of RAVLT factors as the structures more closely reflected those factor loadings than the findings of Donders on the CVLT-II (2008). The two factors of the RAVLT were conceptualized as consolidation/retention and initial learning/recognition; the two factors of the HVLT-R were conceptualized as learning/retention and memory strategy; the three factors of the CVLT-II were conceptualized as retention, initial learning, and memory strategy. The factor structures for all three tasks are illustrated in figures 3-5. Regression formulae were developed from the coefficients derived from the initial analysis and calculated to represent the latent variables identified. These computed factors were then submitted to a second-order exploratory factor analysis to determine the relationship of these factors between measures. The final factor structure contained three factors with the

first including the retention factors from all three tasks, the second including the strategy factors from the CVLT-II and HVLT-R, and the third including the initial learning/storage factor from the RAVLT and the initial learning factor from the CVLT-II. The final model and relationship with the three tasks is illustrated in figure 6.

Discussion

The purpose of the current study was to develop a comprehensive guide exploring the intra-individual variations in performance across three commonly employed VLTs. In addition, the study attempted to explain sources of variance inherent to the structure and construct definitions of each task, and as an additional benefit of the study design, be able to offer updated, rigorously standardized, normative data for the young adult age range in the two VLTs with limited young adult normative data. In review of the results of the current study and to coherently address each study objective, this section will restate the objectives and then discuss the results relevant to addressing those objectives in turn. There was no evidence presented that the member participants in each of the three study cells differed along any primary demographic characteristic. The cells were therefore deemed appropriate for comparison of order effect on VLT performance.

Preliminary Analyses

1. In order to collapse the three groups into a large normative dataset, it is necessary to test for order effects in primary scoring variables (ex: trial 1 total, last trial total, total learning, learning curve, delayed recall). Therefore, initial analyses will target order effects between the three orders of administration.

No order effects were demonstrated with regard to primary scoring variables, including trial learning, overall learning, learning curve, or delayed recall variables. This lack of order effects suggests that neither rote practice nor meta-memoric strategies carried over from one testing session to another in terms of improvement that would affect learning or retention on any measure, maintaining the novelty and integrity of the VLT to assess

verbal learning on repeated testing given time in between (Brandt & Benedict, 1998; Crawford et al., 1989; Delis et al., 2000).

2. Secondary order effects analysis will consider if there are order effects for less commonly used scoring variables (ex: intrusions, clustering strategies, recognition). While these would not preclude the use of the collapsed dataset as a normative base for scoring purposes, it would suggest that caution must be exercised as expanded data may be corrupted by order effects.

Only one intrusion variable demonstrated the expected difference due to order effects. Additionally when the effect was investigated, it demonstrated a paradoxical and likely spurious effect through which final test position was associated with the hypothesized higher rate of intrusions over both the first and middle positions, but middle test position was associated with significantly fewer intrusions over both the final *and* first positions. The only possible explanation available from the current study design is that due to the semi-counterbalance of the Latin Squares Design, the meta-memoric effects and/or memory carryover of the more similar semantic categories of the HVLT-R and CVLT-II resulted in the increased intrusion rate on the CVLT-II in final position as the HVLT-R only precedes the CVLT-II in this position (Group HRC; see appendix 4). Conversely, the practice effects of a previous recognition task, but not proactive interference for semantic intrusions presented by the RAVLT may have led to improved performance in intrusion rates in both the HRC and RCH groups. Alternatively, the effect may simply be a spurious statistical spike and will require replication and possibly full counterbalance design to explore fully. As mentioned previously, the positively skewed distribution of the intrusion variables may be

generating questionable findings and the increased intrusion rate may be an artifact of the floor effect for intrusions on all three tasks.

3. Resulting means and standard deviations for each measure, broken up by gender, will be provided along with demographic information in each cell to provide additional normative data for future study and clinical use of these measures.

The data presented in appendices 9-11 provide a normative dataset for use in the three measures broken by gender, with demographic characteristics to clarify group membership. These data can be used in future studies of young adult populations for either clinical or research purposes and were rigorously standardized in administration as to easily be merged into a meta-analytic study of demographic characteristics and task performance in order to generate more robust normative data in the 18-29 year old age range.

Primary Objectives

1. Comparable scores from each of the three measures will be compared via Pearson correlations. Expectations are that most variables will correlate significantly, consistent with the similar task construction of the three measures, but that correlations will fall too low to consider the measures as alternate forms of one another and raise questions as to sources of variation between such similar tasks.

As expected, overall learning, immediate recall, and delayed recall were correlated across all measures. Additionally, as hypothesized, these correlations did not come near the recommended but liberal alternate forms reliability coefficient of 0.7. However, learning curve variables were not correlated across tests nor were trial correct variables for some trials across some measures. This inconsistency is most easily explained by the underlying

learning factors associated with test characteristics rather than test session variance due to the significant correlations between the CVLT-II and RAVLT. Overall, the findings indicate that the variables under investigation are similar, but by no means statistically comparable between measures.

2. Trend analysis will be employed to explore the similarities or lack thereof between the three measures across learning trials.

The first three trials between the tasks elucidated the significant deviation of the HVLT-R performance within subjects for a 12 member, categorical collection of data. While the CVLT-II and RAVLT demonstrate similar learning curves across all five trials, the HVLT-R deviates immediately and demonstrates fundamentally different learning gains between each of its trials when compared to the CVLT-II and RAVLT. Consequently, it is unclear what additional trials might demonstrate in learning characteristics for participants. That is, the difference between the learning curves of measures is so great that a full five trial HVLT-R curve could not be predicted from performance on the CVLT-II or RAVLT and first three trials of the HVLT-R given the current study design. However, on observation, there is evidence of a plateauing effect between the second and third trials of the HVLT-R which is likely a consequence of ceiling effects for trial three.

3. The serial position effect will be calculated for each of the three measures to see if task characteristics have an effect on this generally robust feature of list learning measures.

The CVLT-II and RAVLT demonstrated typical serial position effects. That is, those items near the beginning and end of the lists were more likely to be recalled those in the

center. However, as figure 2 illustrates, the HVLT-R does not demonstrate a recency effect, but does demonstrate a robust primacy effect. This leaves several possible explanations as to the variation in memory structure on the HVLT-R versus the CVLT-II and RAVLT. As hypothesized above, this could simply be a consequence of a shorter list, resulting in performance high enough to wash away typical serial position effects. However, in the presence of a primacy effect, but no recency effect, there are two related possibilities that must be considered. First, due to the short length, participants may be relying on pure primacy strategies (e.g., rote memorization and rehearsal during list presentation) to learn the list, which may be a more efficient approach given the list length. Similarly, the longer delay between item presentations may pull for this form of strategy given the extra time to review items in presentation format. Either way, the loss of recency effects may be relevant when comparing memory groups for which significant recency effect differences in serial position effects may differentiate memory profiles, such as is the case in Alzheimer's disease versus geriatric depression (Foldi et al., 2003). Therefore, when selecting a VLT, consideration of the referral question may be valuable in determining the best selection of tool. In particular, when considering normal aging versus geriatric depression versus Alzheimer's disease, selection of a tool with more robust pull for serial position effects may better differentiate groups.

Supplemental Objectives

1. Serial clustering effects for the three measures will be evaluated with the expectation that greater serial clustering will be seen overall on the RAVLT relative to the CVLT-II, due to the lack of cuing or overt categories for list items.

The presence of higher RAVLT serial clustering strategies over the other measures indicates that the test characteristics of the RAVLT are such that options for alternative strategies are more limited. Additionally, as the study was performed within subjects, this difference indicates that individual strategies may vary between tasks based on test characteristics. That is, the encoding strategy of an individual may vary with the test they are presented.

2. Similarly, the semantic clustering variable on the CVLT-II is expected to explain differences in serial clustering strategies between the CVLT-II and RAVLT, indicating that one strategy is generally independent from another.

The linear effect of semantic clustering on differences in serial clustering scores indicates an important finding relevant to selection of VLTs, particularly with regard to the primary differences between the CVLT-II and the RAVLT on participant memory strategy. First, due to the significant range of scores on the CVLT-II, some individuals will utilize semantic clustering strategies while others will opt to rely on serial clustering strategies when both options are present. This implies that the approach to learning will differ from the CVLT-II to the RAVLT in some participants who would structure their clustering around semantic clustering when available, but not for others who would continue to attempt a serial memory strategy. Therefore, individual differences in memory strategy will vary with the test characteristics.

3. Expanding this relationship, the use of semantic clustering on the CVLT-II is expected to explain differences in overall performance between the CVLT-II and RAVLT as those who rely on serial clustering strategies will perform similarly on

both tasks, while those who prefer semantic clustering will perform better on the CVLT-II than on the RAVLT.

The effect that semantic clustering has on differences in performance across measures highlights the difficulty in selection of a VLT when considering several options. An individual's performance, when compared to a normative sample, may vary due to intra-individual variations in preferred memory strategy. The implication of the relationship between semantic clustering and differences in a measure with and without semantically related items indicates that there may be serial clusterers and semantic clusterers who identify with one approach or another based on efficiency of the participant's own cognitive ability structure. This is particularly concerning for task selection as the neuropsychologist must now consider whether the verbal learning construct of interest includes meta-memoric semantic clustering strategies or rote serial clustering strategies as key to the memory construct. Further, in the presence of only one administration, it is unclear if an individual would have performed better or worse on the alternative task.

4. With the addition of more composite variables derived from existing scores, factor analysis of the variables from the three tasks is expected to load variables similarly in a factor analysis of the tested composites, forming four factors (Attention Span, Learning Efficiency, Delayed Memory, and Inaccurate Memory) for all three tasks, as seen in earlier study of CVLT-II in normal controls (Donders, 2008).

The results of the second order factor analysis indicate several important characteristics about the VLT testing paradigm. First, and directly relevant to the above unsupported hypothesis, authors have generated variable factor structures for several VLTs, particularly with regard to those that carry more complex procedures, such as the CVLT-II (Delis et al.,

2000; Donders, 2008; Vakil & Blachstein, 1993). The multiple and highly correlated individual variables and relevant composite scores are clearly tapping different constructs, but the interpretation of those constructs continues to be unclear in the absence of a standardized definition of verbal learning and memory as a construct of interest. Therefore, any factor analysis of the structure of a given task is likely to generate multiple interpretations depending on the author's choice of how to include variables and composites of interest (Vakil & Blachstein, 1993). However, by applying the results of factor analysis to a higher order analysis of structure, one can identify that there is shared variance across all three VLTs of interest and that, to some extent, they are all tapping similar constructs of interest.

Limitations

Despite the relevant findings of the current study, several limitations need to be addressed in drawing generalizable conclusions from these results. First, the lack of clinical populations being represented is notable as the factor structure of a VLT is likely to change with the neurological implications of the group in question (Delis et al., 2000). As argued elsewhere, due to the complexity of memory deficit profiles, mixed clinical groups are insufficient to identify variations in factor structure within groups, limiting both clinical validity of test results for an individual and research gains in understanding neurocognitive pathway relationships in verbal memory. Therefore, replication of the current study with well-defined neurological populations of interest is highly recommended to provide clinically meaningful and research relevant information regarding profile differences across neurologically impaired populations. Despite this, the current study reaches its prescribed

goal of providing early information from a normal control population for the justification of need in analyzing more neurologically complex groups.

Another limitation for generalization of findings for the current study is the restricted age range. While the age range presented provides some benefit for the establishment of more solid young adult normative data, it also limits the generalizability of the study findings across relevant age bands where common clinical memory concerns may need to be addressed via repeated verbal learning assessment (i.e., dementia). It has also been demonstrated that age plays a factor in the effect on performance of at least one variable of interest from the current study (Meijer et al., 2006). Therefore, this is an area for future assessment along the developmental trajectory, using the current study as a starting point for development.

Motivation was assessed only at the third testing point and introduces the possibility that motivation may have reached significantly low levels on either the first or second administration. However, due to the need for novelty in the task, it was only feasible to administer the symptom validity task once. To address this qualitatively, participants were observed and logged by examiners for behaviors consistent with poor motivation and reviewed with either the research coordinator or the principal investigator.

Finally, while careful steps were taken to ensure rigorous standardized administration and scoring guidelines, there were several participants lost to attrition and administrative error. Associated with this limitation is the somewhat restricted sample size for several of the more power-intensive analyses employed, including factor analysis and formation of normative reference for an age-group of interest. While the current sample is sufficient for

the analyses included, more complex factor analyses of the task variables will require a larger sample to be collected.

Conclusions

There are several conclusions that can be drawn from the current study by focusing on specific hypotheses and the differences illustrated by intra-individual variations in performance. First, while there are many similarities among the three VLTs under investigation, the key differences are significant and result in low coefficients across tasks for conceptually similar variables. The coefficient levels indicate two important consequences of variations in the testing paradigm: low but significant coefficients make estimation of gains/losses or extrapolation of information from one test into the context of another untenable approaches to interpretation of test results; however, the presence of correlations between so many similar variables also indicates that a similar memory construct is being tapped by all three measures despite the variations in task parameters and as such no test can be clearly defined as the "best verbal learning test" due to the lack of a formalized definition of the construct of interest. Current recommendations would include the use of the HVLT-R in those situations in which repeatability, rapid fatigue, and a lowered floor are of importance. Therefore, dementing populations are still the primary target group for this test. When concerns are primarily in the area of consolidation of material or level of memory impairment, it may be more beneficial to rely on the RAVLT, which excludes the effect of semantic categorization, an effect which may elevate performance in some populations (Vanderploeg, Crowell, & Curtiss, 2001). However, there is also significant value in the use of measuring verbal learning in the presence of semantic categorization and cueing components which may provide a more ecologically valid approach to verbal memory when

developing treatment plans and measuring intervention outcomes (DeLuca & Beers, 2004). In short, consideration of the task parameters and construct of interest for the purpose of assessment must be considered in choosing the best VLT.

Related to this recommendation, the second major conclusion of the current study is that in defining the construct of verbal learning for a given study or patient population, the neuropsychologist must consider the purpose for evaluation of verbal learning and select a test that best meets the needed task. Of particular importance is the relationship between "pure" rote verbal memory and strategic meta-memory analysis. The presence of semantic clustering structure in a VLT may or may not alter the approach taken by a particular individual. As a consequence, in the setting of a single evaluation, the neuropsychologist must choose whether the patient is best served by data on rote list memory, which may be helpful in clarifying reduced memory span, or selection of a measure with semantic structure, which may provide a more ecologically valid estimate of everyday verbal memory function. The answer will vary with the evaluation, but the need for both measures is clear in both the clinical and research arenas. One possibility would be the inclusion of a separate verbal learning measure that would look at only serial information and then use a semantic task as well and calculate variability between the two.

Third, the process of learning as illustrated by the learning curves of the three measures, is clearly more complex than is easily estimated by the linear calculation of learning curve as best trial minus first trial. Additionally, the test characteristics of measures are clearly associated with changes in acquisition of information, particularly with regard to the difference between CVLT-II and RAVLT performance versus HVLT-R performance. The complexity of this interaction when considering the performance of neurologically

impaired individuals cannot be overstated. Additional study of learning curves on VLTs and acquisition of verbal information must be better understood before broad statements can be made as to the meaning of variations in the learning curve. As it stands, the only firm statement that can be made with regards to learning curves is that significant failure to make any gains across multiple trials is indicative of significant encoding deficits. Moreover, the lack of correlation between learning curves within subjects across all measures seems to suggest that there may not be a clear construct of "verbal learning rate" and that performance is too highly affected by individual differences and test characteristics to be useful in clinical analysis of subtle memory deficits.

Finally, factor structure of the verbal learning paradigm is an ongoing process with no clear factor structure arising as definitive. This is likely a consequence of differences in interpretation as to what variables should be considered relevant to the structure of the verbal learning and memory construct. However, the similarities in meta-factor loadings for the three measures clearly indicate that the measures share enough common features under the VLT testing paradigm to produce reasonable estimates of learning and retention. However, the other test characteristics may alter this structure with clustering strategies which have been emphasized in the findings above as a primary source of variance in performance within participants and in the absence of firm acknowledgement or field-wide concordance with regards to construct definition, selection of a verbal learning construct that includes or excludes categorical memory strategy will remain at the clinical judgment of the neuropsychologist. This is not to imply that the selection is arbitrary, but much to the contrary, that it is vital that the construct of interest be clearly defined in the

neuropsychologist's mind prior to test selection and access to multiple measures may be necessary in a practice that receives various referral questions.

There are also implications of these findings that are relevant for both practicing clinical neuropsychologists and researchers engaging in translational research in verbal memory. Due to the variability in performance across measures in normal controls, statements commonly employed in clinical literature, such as a statement that a disease process affects performance on verbal learning in which the construct of interest is stated, but not defined, should be modified to include both a definition of the construct and information regarding the specific measure from which this conclusion was drawn to avoid over-generalizing neuropsychological deficits demonstrated on a single measure as reified deficits of the verbal learning construct, which is not clearly defined in the literature.

References

Anderson-Haney, C., Sherman, M. L., Riggs, R., Agocha, V. B., & Compas, B. E. (2003). Neuropsychological effects of treatment for adults with cancer: A meta-analysis and review of the literature. *Journal of the International Neuropsychological Society, 9*, 967-982. doi: 10.1017/S1355617703970019

Belanger, H. G., Curtiss, G., Demery, J. A., Lebowitz, B. K., & Vanderploeg, R. D. (2005). Factors moderating neuropsychological outcomes following mild traumatic brain injury: A meta-analysis. *Journal of the International Neuropsychological Society, 11*, 215-227. doi: 10.1017/S1355617705050277

Belanger, H. G. & Vanderploeg, R. D. (2005). The neuropsychological impact of sports related concussion: A meta-analysis. *Journal of the International Neuropsychological Society, 11*, 345-347. doi: 10.1017/S1355617705050411

Benedict, R. H. B., Schretlen, D., Groninger, L., & Brandt, J. (1998). Hopkins Verbal Learning Test – Revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist, 12*, 43-55. doi: 10.1076/clin.12.1.43.1726

Boake, C. (2002). Edouard Claparede and the auditory verbal learning test. *Journal of Clinical and Experimental Neuropsychology, 22,* 286-292. doi: 10.1076/1380-3395(200004)22:2;1-1;FT286

Brandt, J. & Benedict, R. H. B. (2007). *Hopkins Verbal Learning Test – Revised (HVLT-R).* Psychological Assessment Resources.

Campbell, L. K., Scaduto, M., Sharp, W., Dufton, L., Van Slyke, D., Whitlock, J. A., &

Compas, B. (2007). A meta-analysis of the neurocognitive sequelae of treatment for childhood acute lymphocytic leukemia. *Pediatric Blood & Cancer, 49,* 65-73. doi: 10.1002/pbc.20860

Crawford, J. R., Stewart, L. E., & Moore, J. W. (1989). Demonstration of savings on the AVLT and development of a parallel form. *Journal of Clinical and Experimental Neuropsychology, 11,* 975-981. doi: 10.1080/01688638908400950

Crossen, J. R. & Wiens, A. N. (1994). Comparison of the Auditory-Verbal Learning Test (AVLT) and California Verbal Learning Test (CVLT) in a sample of normal subjects. *Journal of Clinical and Experimental Neuropsychology, 16,* 190-194. doi: 10.1080/01688639408402630

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,* 754-768. doi: 10.1080/01688638808402812

Delaney, R. C., Prevey, M. L., Cramer, J., Mattson, R. H., et al. Test-retest comparability and control subject data for the Rey-Auditory Verbal Learning Test and Rey-Osterrieth/Taylor Complex Figures. *Archives of Clinical Neuropsychology, 7,* 523-528. doi: 10.1016/0887-6177(92)90142-A

Delis, D. C., Freeland, J., Kramer, J. H., & Kaplan, E. (1988). Integrating clinical assessment with cognitive neuroscience: Construct validation of the California Verbal Learning Test. *Journal of Consulting and Clinical Psychology, 56,* 123-130. doi: 10.1037/0022-006X.56.1.123

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *California Verbal Learning*

Test – Second Edition (CVLT-II). Pearson Assessments.

Delis, D. C., Massman, P. J., Butters, N., Salmon, D. P., Cermak, L. S., Kramer, J. H. (1991). Profiles of demented and amnesic patients on the California Verbal Learning Test: Implications for the assessment of memory disorders. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*, 19-26. doi: 10.1037/1040-3590.3.1.19

DeLuca, J. & Chiaravalloti, N. D. (2004). Memory and learning in adults. In G. Goldstein & S. R. Beers (Eds), *Comprehensive Handbook of Psychological Assessment Volume 1: Intellectual and Neuropsychological Assessment (pp. 217-236).* Hoboken, New Jersey: John Wiley & Sons Inc.

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. doi: 10.1177/1073191107310926

Drummond, S. P. A., Brown, G. G., Gillin, J. C., Stricker, J. L., Wong, E. C., & Buxton, R. B. (2000). Altered brain response to verbal learning following sleep deprivation. *Nature, 403*, 655-657. doi: 10.1038/35001068

Foldi, N. S., Brickman, A. M., Schaefer, L. A., & Knutelska, M. E. (2003). Distinct serial position profiles and neuropsychological measures differentiate late life depression from normal aging and Alzheimer's disease. *Psychiatry Research, 120*, 71-84. doi: 10.1016/S0165-1781(03)00163-X

Geffen, G. M., 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. *Clinical Neuropsychologist, 4,* 45-63. doi: 10.1080/13854049008401496

Hiscock, M. & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11,* 967-974. doi: 10.1080/01688638908400949

Lacritz, L. H., Cullum, C. M., Weiner, M. F., & Rosenberg, R. N. (2001). Comparison of the Hopkins Verbal Learning Test – Revised to the California Verbal Learning Test in Alzheimer's Disease. *Applied Neuropsychology, 8,* 180-184. doi: 10.1207/S15324826AN0803_8

Lacritz, L. H. & Cullum, C. M. (1998). The Hopkins Verbal Learning Test and CVLT: A preliminary comparison. *Archives of Clinical Neuropsychology, 13,* 623-628. doi: 10.1016/S0887-6177(98)00004-3

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). Neuropsychological Assessment: Fourth Edition. New York, New York: Oxford University Press.

Loring, D., Lee, G., & Meador, J. (2005). Victoria Symptom Validity Test performance in nonlitigating epilepsy surgery candidates. *Journal of Clinical and Experimental Neuropsychology, 27,* 610-617. doi: 10.1080/01688639608408313

Meijer, W. A., de Groot, R. H. M., van Boxtel, M. P. J., van Gerven, P. W. M., & Jones, J. (2006). Verbal learning and aging: Combined effects of irrelevant speech, interstimulus interval, and education. *Journal of Gerontology, 61B,* 285-294. doi: 10.1093/geronb/61.5.P285

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). Handbook of Normative

Data for Neuropsychological Assessment. New York, New York: Oxford University Press.

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. doi: 10.1016/j.acn.2004.02.005

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives de Psychologie, 28*, 286-340.

Runquist, W. N. (1966). Intralist interference as a function of list length and interstimulus similiarity. *Journal of Verbal Learning and Verbal Behavior, 5,* 7-13. doi: 10.1016/S0022-5371(66)80099-3

Shapiro, A. M., Benedict, R. H. B., Schretlen, D., & Brandt, J. (1999). Construct and concurrent validity of the Hopkins Verbal Learning Test – Revised. *The Clinical Neuropsychologist, 13,* 349-358. doi: 10.1076/clin.13.3.348.1749

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A Handbook.* Los Angeles, California: Western Psychological Services.

Shear, P. K., Wells, C. T., & Brock, A. M. (2000). The effect of semantic cuing on CVLT performance in healthy participants. *Journal of Clinical and Experimental Neuropsychology, 22,* 649-655. doi: 10.1076/1380-3395(200010)22:5;1-9;FT649

Slick, D. J., Hopp, G., Strauss, E., Hunter, M., & Pinch, D. (1994). Detecting dissimulation: Profiles of simulated malingerers, traumatic brain-injury patients, and normal controls on a revised version of Hiscock and Hiscock's forced-choice memory test. *Journal of*

Clinical and Experimental Neuropsychology, *16*, 472-481. doi: 10.1080/01688639408402657

Slick, D. J., Tan, J., Strauss, E., & Mateer, C. (2003). Victoria Symptom Validity Test scores of patients with profound memory impairment: Nonlitigant case studies. *The Clinical Neuropsychologist, 17,* 390-394. doi: 10.1080/01688639608408313

Squire, L. R. & Schacter, D. L. (2003). *Neuropsychology of Memory, Third Edition.* New York: New York. The Guilford Press.

Strauss, E., Sherman, E.M.S., & Spreen, O. (2006). *A Compendium of neuropsychological tests: Administration, norms, and commentary.* New York: New York. Oxford University Press.

Stallings, G., Boake, C. & Sherer, M. (1995). Comparison of the California Verbal Learning Test and the Rey Auditory Verbal Learning Test in Head-Injured Patients. *Journal of Clinical and Experimental Neuropsychology*, *17*, 706-712. doi: 10.1080/01688639508405160

Urbaniak, G. C., & Plous, S. (2011). Research Randomizer (Version 3.0) [Computer software]. Retrieved on April 22, 2011, from http://www.randomizer.org/

Vakil, E. & Blachstein, H. (1993). Rey Auditory-Verbal Learning Test: Structure analysis. *Journal of Clinical Psychology, 49,* 883-890. doi: 10.1002/1097-4679(199311)49:6<883::AID-JCLP2270490616>3.0.CO;2-6

Vanderploeg, R. D., Crowell, T. A., & Curtiss, G. (2001). Verbal learning and memory deficits in traumatic brain injury: Encoding, consolidation, and retrieval. *Journal of Clinical and Experimental Neuropsychology, 23,* 185-195. doi: 10.1076/jcen.23.2.185.1210

van der Elst, W., van Boxtel, M. P. J., van Breukelen, G. J. P., & Jolles, J. (2005). Rey's verbal learning test: Normative data for 1855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11,* 290-302. doi: 10.10170S1355617705050344

Wright, A. A. (1998). Auditory list memory in rhesus monkeys. *Psychological Science, 9,* 91-98. doi: 10.1111/1467-9280.00017

Table 1

Pearson Correlations for VLTs

Measure	CVLT-HVLT	*p*	HVLT-RAVLT	*p*	CVLT-RAVLT	*p*
Trial 1 Correct	0.17	0.11	0.08	0.47	0.31*	0.002
Trial 2 Correct	0.29*	0.001	0.29*	0.005	0.33*	0.001
Trial 3 Correct	0.15	0.16	0.26*	0.01	0.42*	<0.001
Trial 4 Correct	N/A	N/A	N/A	N/A	0.49*	<0.001
Trial 5 Correct	N/A	N/A	N/A	N/A	0.39*	0.001
Total Learning	0.41*	<0.001	0.32*	0.002	0.56*	<0.001
Learning Curve	-0.02	0.86	0.13	0.27	0.12	0.29
Immediate Recall	N/A	N/A	N/A	N/A	0.5*	<0.001
Delayed Recall	0.42*	<0.001	0.41*	<0.001	0.52*	<0.001
True Positives	0.22*	0.04	0.22*	0.04	0.42*	<0.001
False Positives	0.44*	<0.001	0.21*	0.05	0.52*	<0.001

Notes: Correlations between each pair of measures presented.

** indicates significance of correlation at $p \leq 0.05$.*

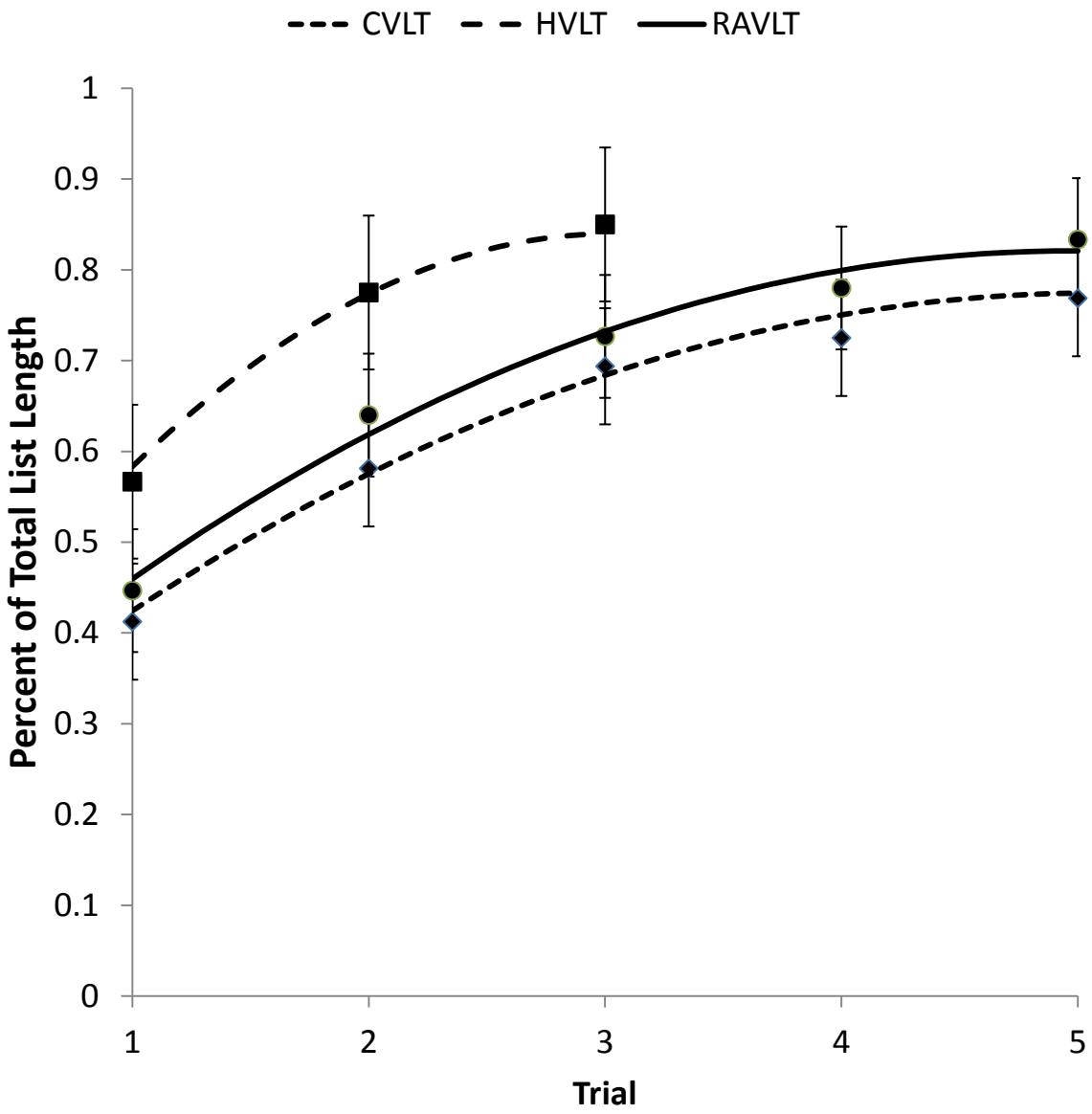


Figure 1. Learning Curves across Tasks. The CVLT-II and RAVLT have 5 trial learning curves, but the HVLT-R can only produce a 3 trial learning curve. Y-axis values indicate percentage of total words recalled (16 on the CVLT-II, 12 on the HVLT-R, and 15 on the RAVLT).

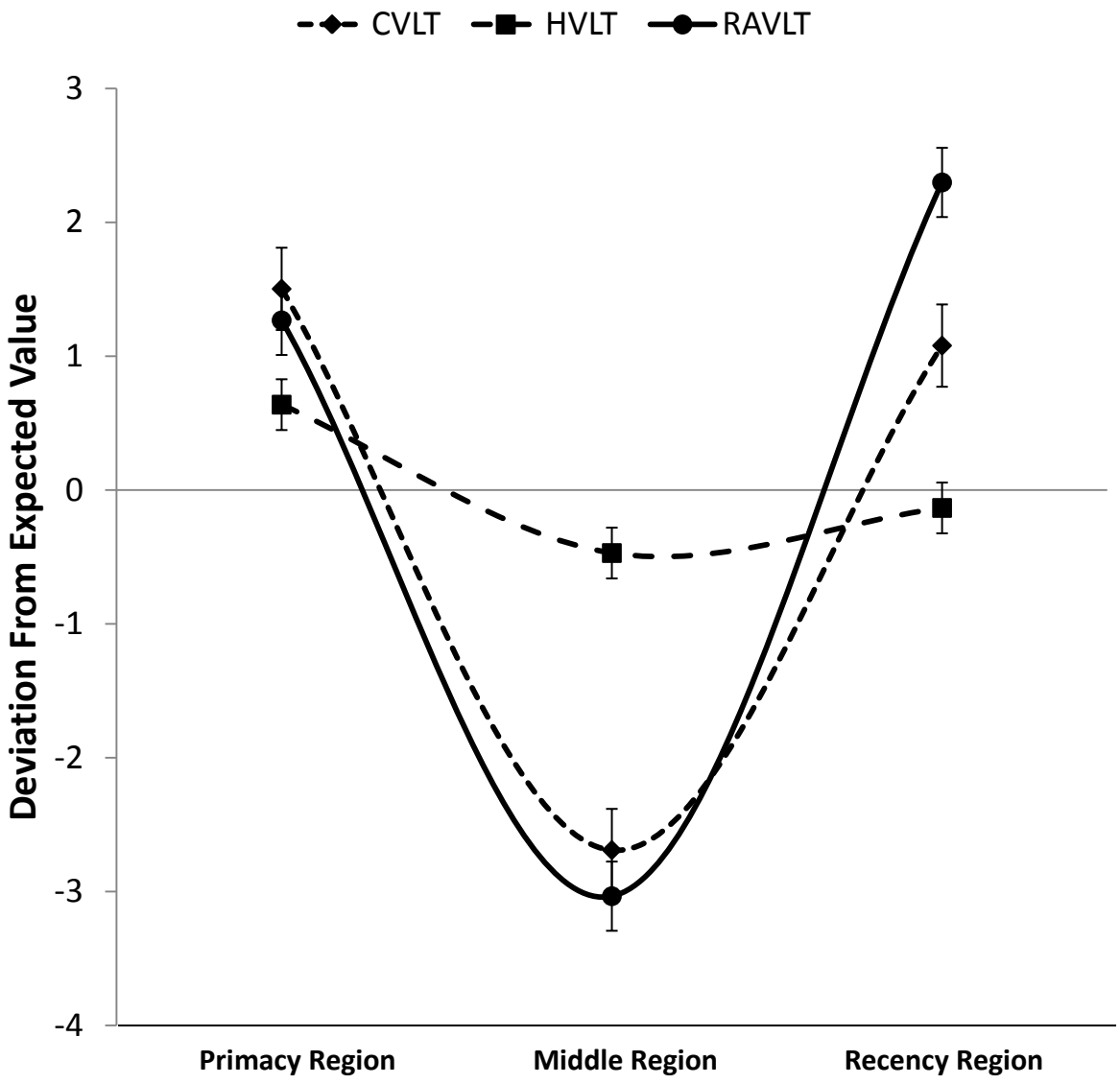


Figure 2. Serial Position Effects by VLT. Serial position effect is indicated with deviation from an expected value if no effect of position is assumed. Therefore, deviations from zero in the positive direction on the primacy and recency regions and deviations from zero in the negative direction on the middle region all indicate the presence of a serial position effect.

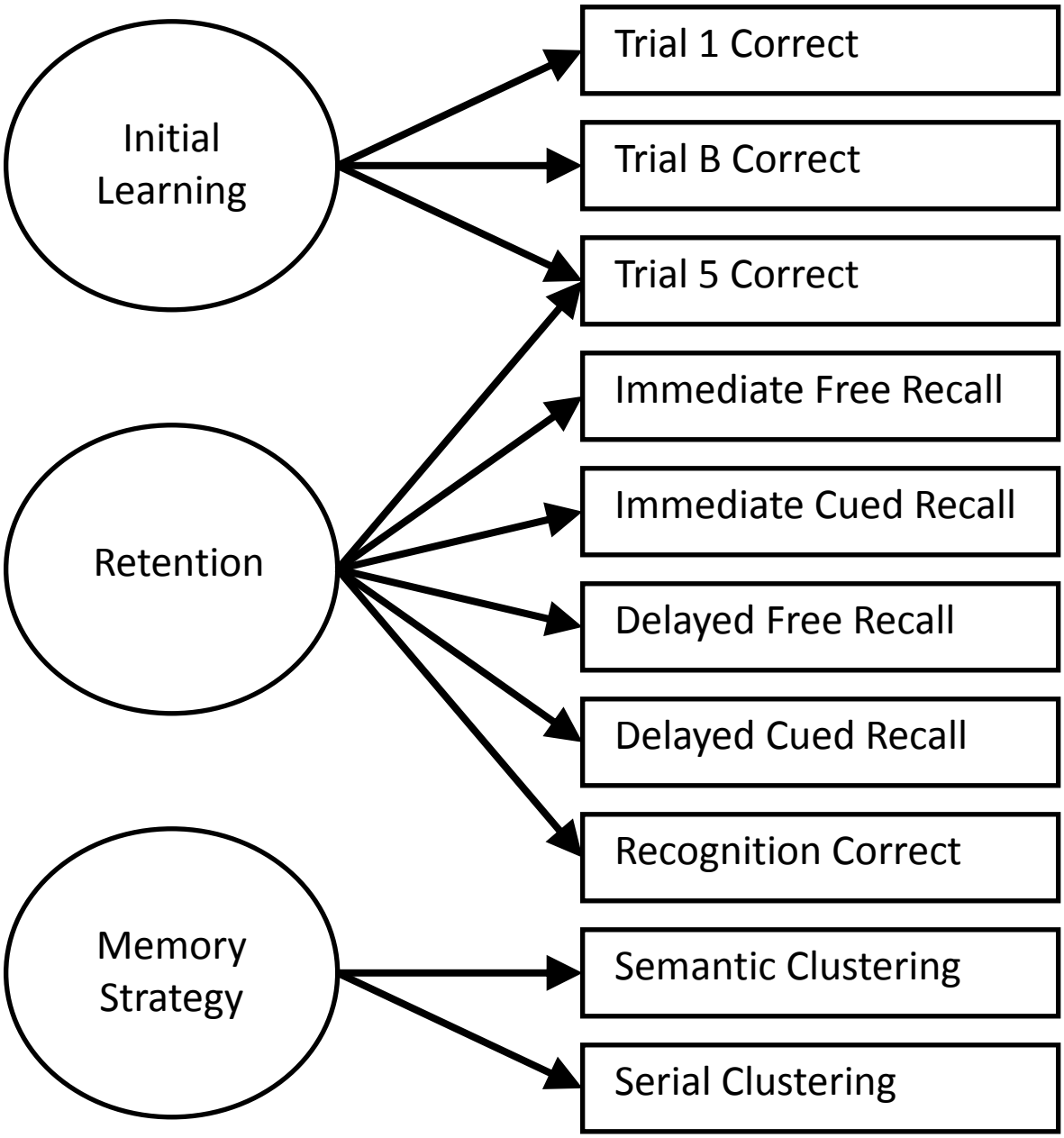


Figure 3. Factor Structure of CVLT-II. Three factor structure of the CVLT-II.

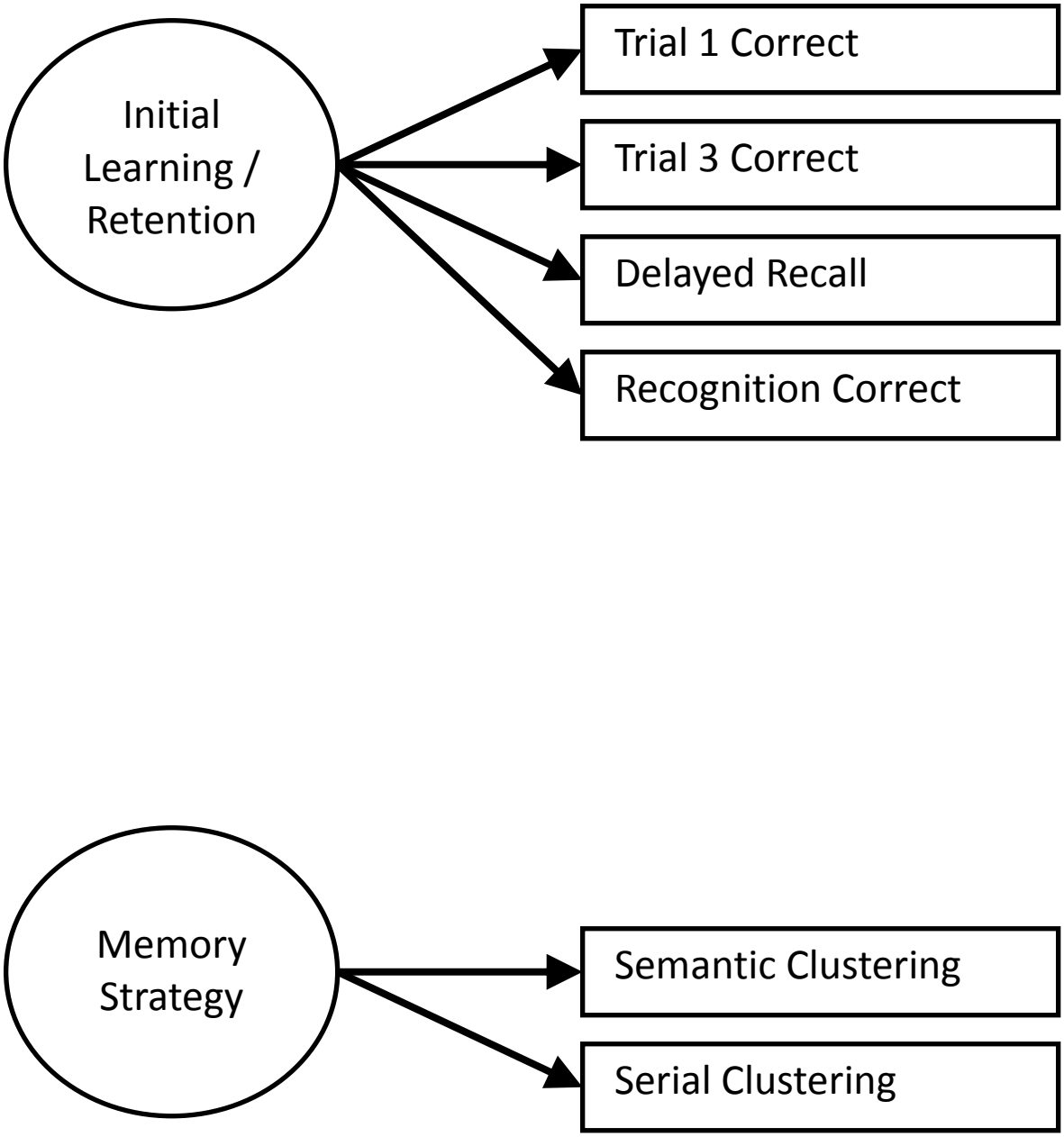


Figure 4. Factor Structure of HVLT-R. Two factor structure of the HVLT-R

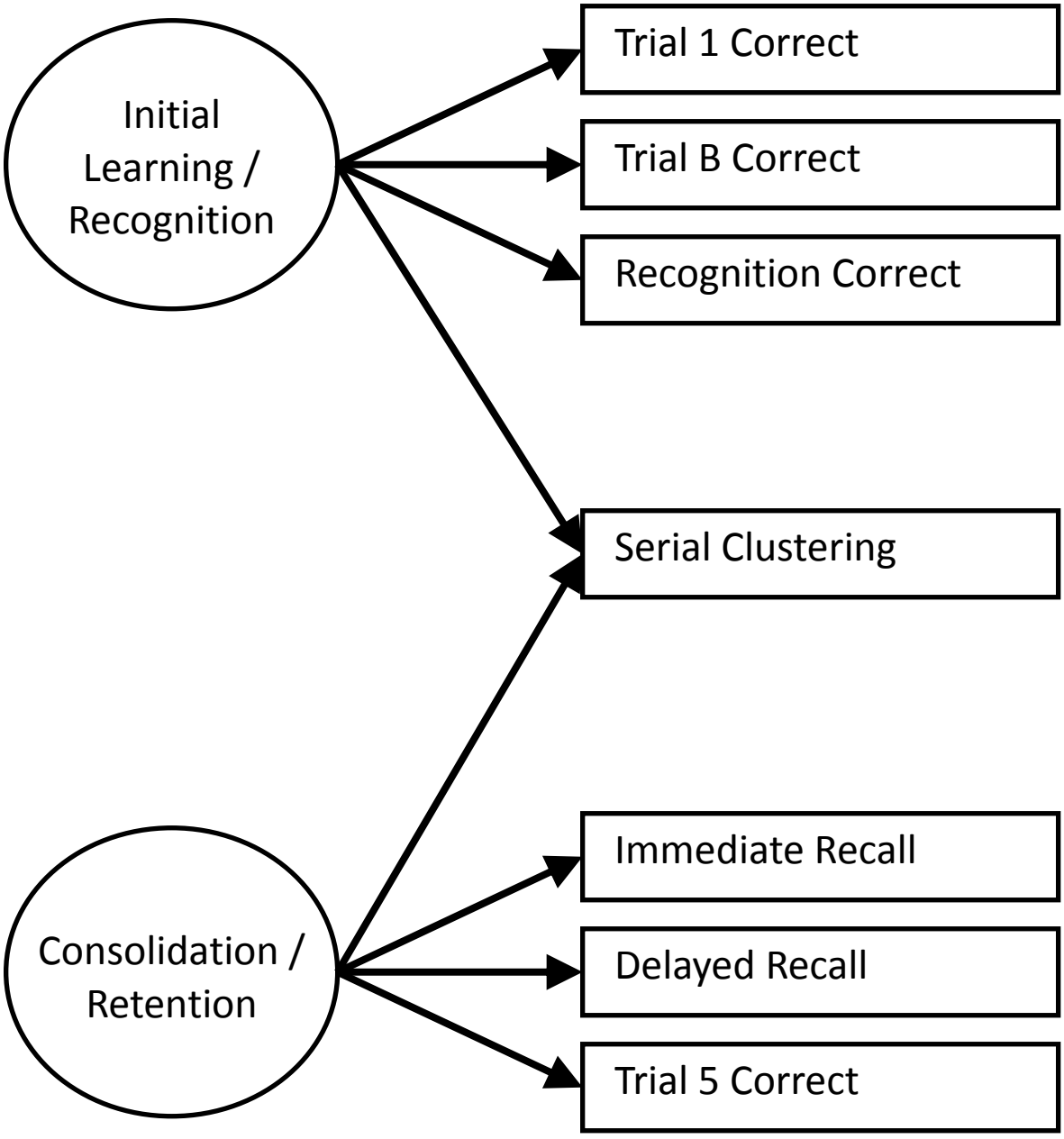


Figure 5. Factor Structure of RAVLT. Two factor structure of the RAVLT.

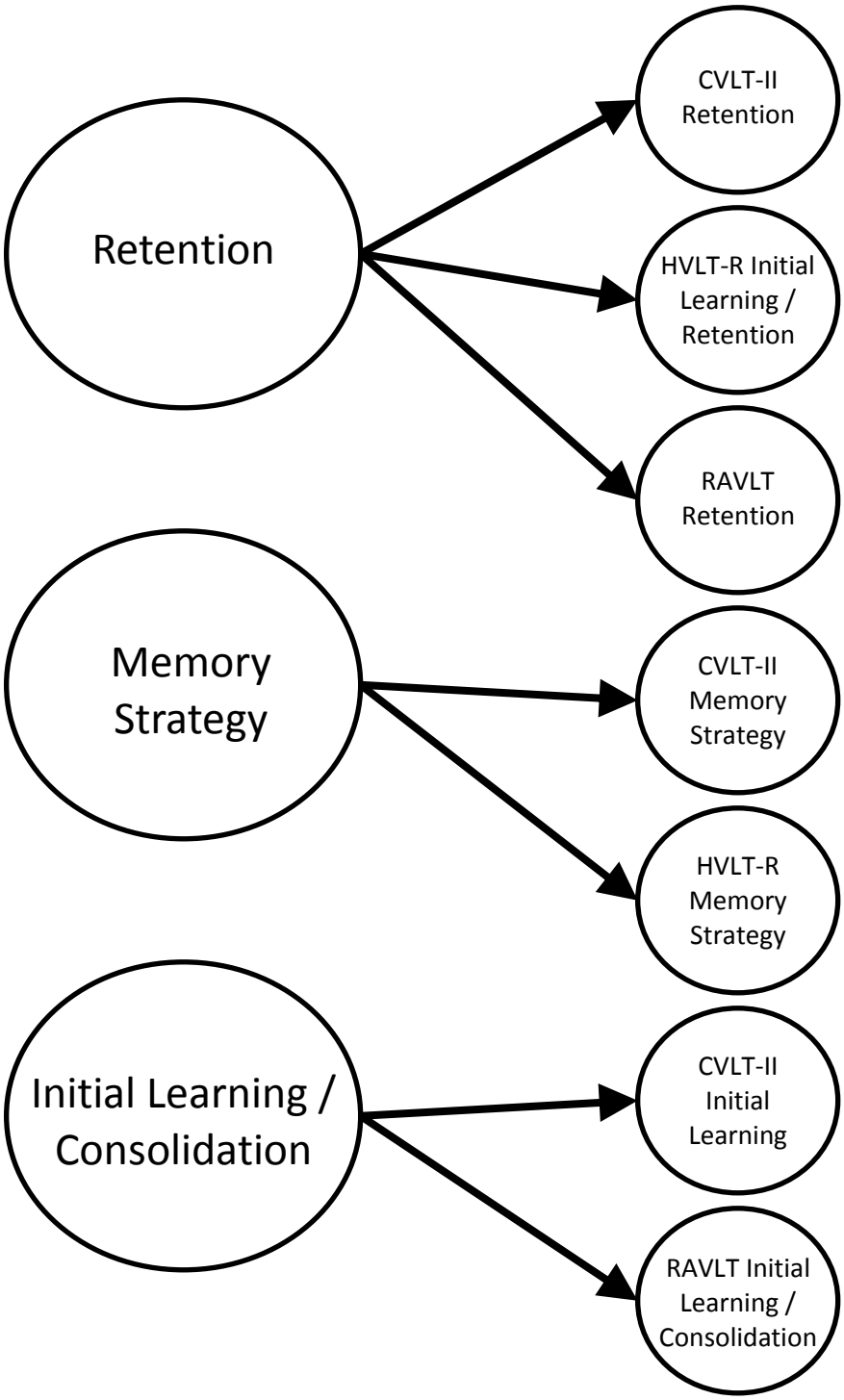
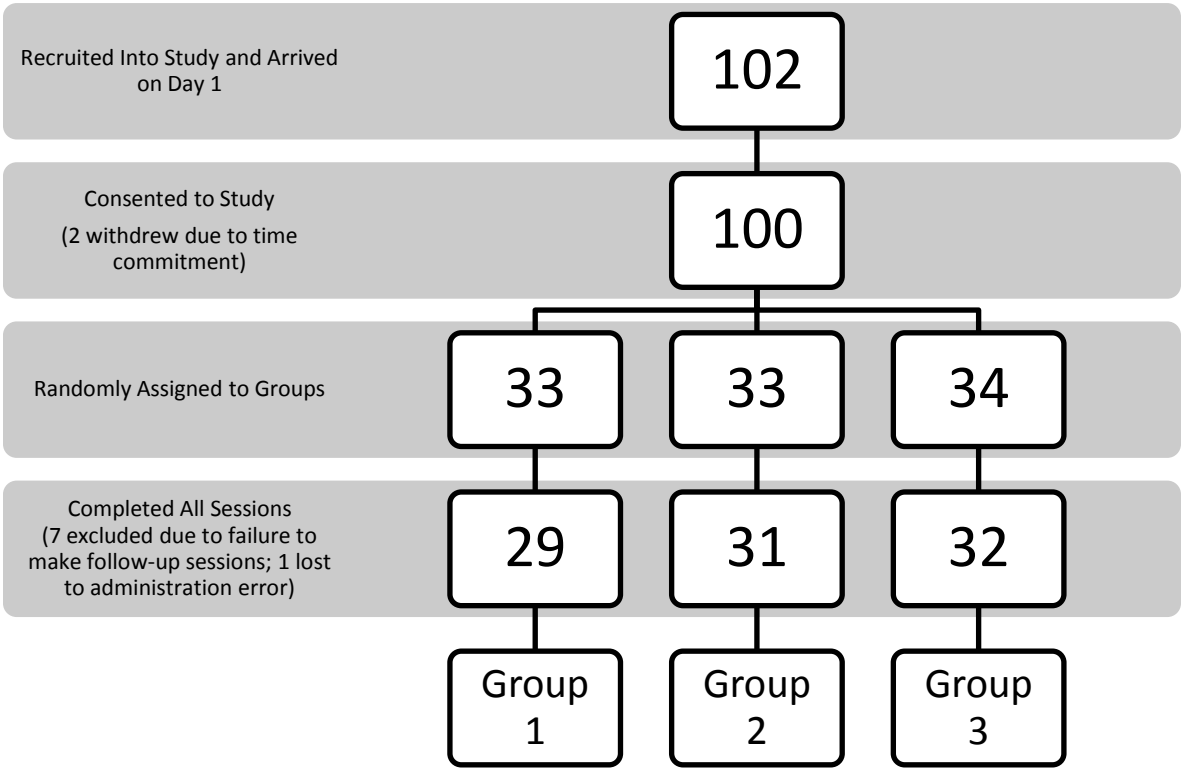


Figure 6. Factor Structure of VLTs from second-order factor analysis.

Appendix 1. Test Characteristics that vary across Tasks.

	RAVLT	**CVLT-II**	**HVLT-R**
Number of Items	15 item list	16 item list	12 item list
Number of Trials	5 learning trials	5 learning trials	3 learning trials
Semantic Categories	N/A	4 categories	3 categories
Free Recall	After each trial	After each trial	After each trial
Interference Trial	Yes	Yes	No
Post-Interference Recall	Yes	Yes	N/A
Cued Recall	No	Yes	No
Delayed Recall	Yes	Yes	Yes
Delay Length	Varies	20 minutes	20-25 minutes
Recognition Trial	Yes	Yes	Yes
Publisher	Public Domain	PsychCorp	PAR
Cost	Free	$499	$247
Normative Data	Multiple Sources	Manualized	Manualized

Appendix 2. Participant Assignment and Attrition.



Appendix 3. Global and Cell Level Demographics.

	Global		CHR		HRC		RCH			
Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD	F(2,89)	*p*
Age	22.48	6.34	22.30	6.41	22.81	7.00	22.32	5.74	0.06	0.94
Years of Education	13.82	1.66	13.90	2.11	13.78	1.60	13.77	1.23	0.05	0.95

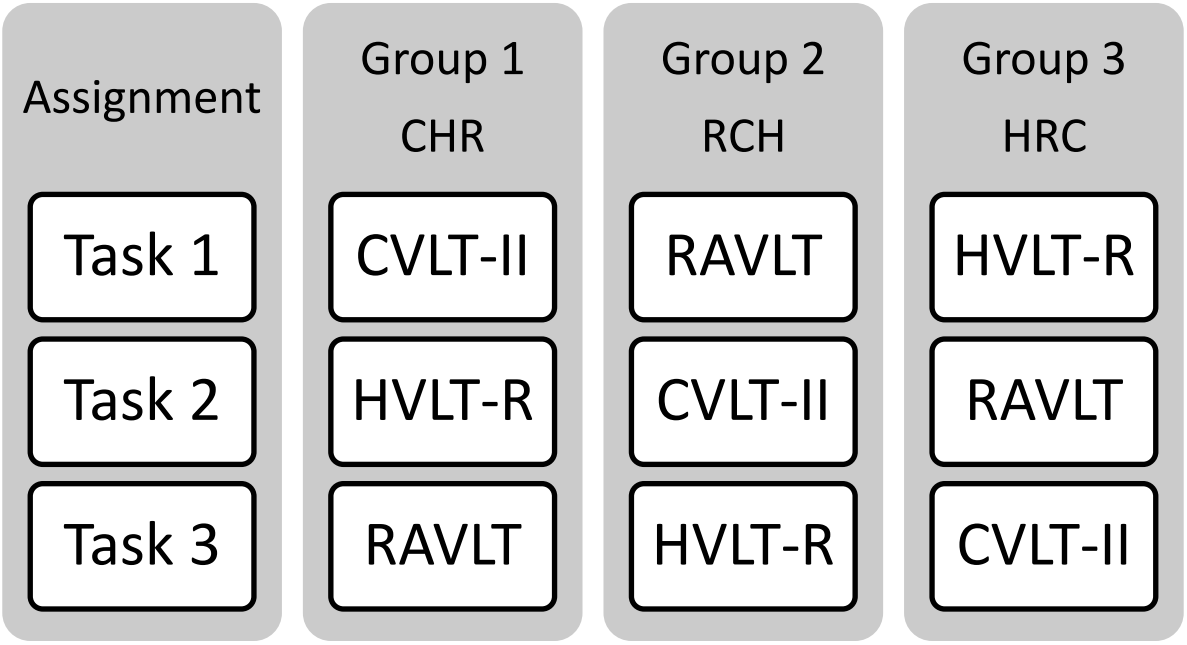
	Global	CHR	HRC	RCH		
Measure	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	χ2(2, N = 89)	*p*
Sex (male)	29.35	34.48	28.13	25.81	0.58	0.75

	Global	CHR	HRC	RCH		
Measure	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	χ2(10, N = 89)	*p*
Race					10.91	0.36
Asian	15.22	3.45	21.88	19.35		
Black	19.57	31.03	12.50	16.13		
Hawaiian Native	2.17	3.45	0.00	3.23		
Native American	2.17	0.00	3.13	3.23		
White	47.83	48.28	43.75	51.61		
Not Reported	13.04	13.79	18.78	6.45		

	Global	CHR	HRC	RCH		
Measure	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	χ2(4, N = 89)	*p*
Ethnicity					4.45	0.35
Hispanic	32.61	24.14	43.75	29.03		
Not Hispanic	53.26	55.17	50.00	54.84		
Not Reported	14.13	20.69	6.25	16.13		

Note: Three letter alphabetisms summarize order of tasks (C = CVLT-II, H = HVLT-R, R = RAVLT). No significant differences were identified.

Appendix 4. Latin Squares Design.



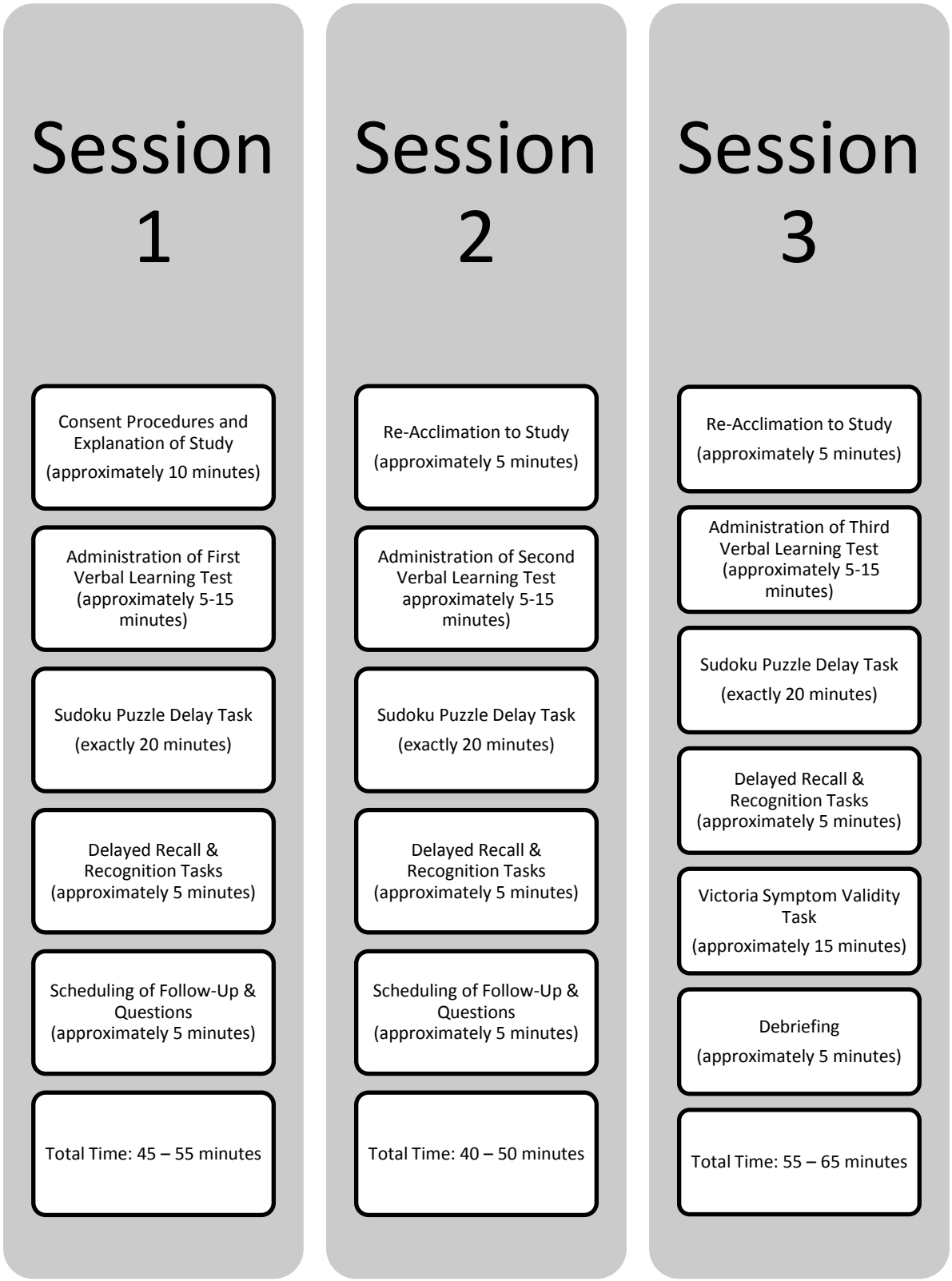
Note: Three letter alphabetisms summarize order of tasks (C = CVLT-II, H = HVLT-R, R = RAVLT).

Appendix 5. Table of Random Assignment of Participants.

Research Randomizer Results:			
40 Sets of 3 Unique Numbers Per Set			
Range: From 1 to 3 -- Unsorted			
Set 1	**Set 11**	**Set 21**	**Set 31**
1	3	2	2
3	2	3	1
2	1	1	3
Set 2	**Set 12**	**Set 22**	**Set 32**
1	3	1	3
2	2	3	1
3	1	2	2
Set 3	**Set 13**	**Set 23**	**Set 33**
1	2	1	2
3	1	2	1
2	3	3	3
Set 4	**Set 14**	**Set 24**	**Set 34**
1	1	2	3
2	2	3	*1*
3	3	1	*2*
Set 5	**Set 15**	**Set 25**	***Set 35***
3	3	1	*3*
2	2	3	*2*
1	1	2	*1*
Set 6	**Set 16**	**Set 26**	***Set 36***
3	2	1	*2*
2	3	2	*1*
1	1	3	*3*
Set 7	**Set 17**	**Set 27**	***Set 37***
3	1	2	*2*
1	2	1	*3*
2	3	3	*1*
Set 8	**Set 18**	**Set 28**	***Set 38***
2	3	1	*2*
1	2	2	*3*
3	1	3	*1*
Set 9	**Set 19**	**Set 29**	***Set 39***
2	3	1	*3*
1	1	3	*1*
3	2	2	*2*
Set 10	**Set 20**	**Set 30**	***Set 40***
2	3	1	*1*
3	2	3	*3*
1	1	2	*2*

Note: Italicized and underlined values were generated but not assigned due to sample size.

Appendix 6. Sample Session Format.



Appendix 7. Variables under investigation in the current study.

C	Calculated by default for the given test
S	Can be calculated using a synonymous formula, but does not have normative data
X	Cannot be calculated for the test
N	Novel variable intended to correct for some characteristic variability

Variable	Calculation	CVLT-II	HVLT-R	RAVLT
Trial 1 Correct	Sum of correct free recalls on trial 1	C	S	C
Trial 2 Correct	Sum of correct free recalls on trial 2	C	S	C
Trial 3 Correct	Sum of correct free recalls on trial 3	C	S	C
Trial 4 Correct	Sum of correct free recalls on trial 4	C	X	C
Trial 5 Correct	Sum of correct free recalls on trial 5	C	X	C
Total 5 Trial Learning	Sum of correct recalls on all 5 trials	C	C	C
3 Trial Learning Curve	Difference of trial 1 from trial 3 hits	N	N	N
5 Trial Learning Curve	Difference of trial 1 from trial 5 hits	N	X	C
Semantic Clustering Score	Number of items on trials 1 - 5 with semantic grouping	C	S	X
Total Immediate Intrusions	Number of intrusion items on trials 1 - 5	C	S	S
Delayed Recall Hits	Number of correct free recalls on delay trial	C	C	C
Delayed Recall Intrusions	Number of intrusion items on delayed recall	C	S	C
Retention	Delayed recall hits divided by trial 5 hits	C	C	S

Appendix 8. Order Effects.

	Global		CHR		HRC		RCH			
Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD	F(2,89)	*p*
CVLT Trial 1 Correct	6.76	1.62	6.72	1.67	6.84	1.95	6.71	1.19	0.06	0.94
HVLT Trial 1 Correct	7.00	1.66	6.97	2.10	6.94	1.32	7.10	1.58	0.08	0.92
RAVLT Trial 1 Correct	6.77	1.54	6.55	1.48	6.78	1.54	6.97	1.62	0.54	0.58
CVLT Total Learning	51.34	8.26	51.45	9.43	51.50	8.52	51.06	6.99	0.03	0.98
HVLT Total Learning	26.36	3.86	26.45	4.38	25.72	3.54	26.94	3.70	0.79	0.46
RAVLT Total Learning	51.45	7.53	50.97	8.13	51.16	7.67	52.23	6.96	0.24	0.78
CVLT Immediate Recall	10.98	2.82	11.10	3.03	10.69	2.80	11.16	2.70	0.26	0.77
RAVLT Immediate Recall	10.86	2.76	10.93	3.22	10.47	2.71	11.19	2.37	0.55	0.58
CVLT Delayed Recall	11.30	2.46	11.21	2.29	11.00	2.46	11.71	2.62	0.69	0.51
HVLT Delayed Recall	9.27	1.75	9.17	1.75	9.34	1.49	9.29	2.02	0.07	0.93
RAVLT Delayed Recall	10.74	2.96	10.14	3.48	10.47	2.88	11.58	2.36	2.02	0.14
CVLT Total Intrusions	1.16	1.81	1.31	1.95	1.53	2.17	0.65	1.05	2.07	0.13
HVLT Total Intrusions	0.49	0.9	0.34	0.67	0.53	0.72	0.58	1.20	0.57	0.57
RAVLT Total Intrusions	0.84	1.48	1.03	1.80	1.09	1.44	0.39	1.05	2.24	0.11

CVLT Intrusions Immediate Recall	0.26	0.51	0.31	0.54	0.41	0.61	0.06	0.25	3.97	0.02*
RAVLT Intrusions Immediate Recall	0.22	0.41	0.28	0.45	0.28	0.46	0.10	0.30	2.02	0.14
CVLT Intrusions Delayed Recall	0.41	0.63	0.41	0.68	0.59	0.67	0.23	0.50	2.78	0.07
HVLT Intrusions Delayed Recall	0.24	0.72	0.17	0.47	0.22	0.91	0.32	0.70	0.34	0.71
RAVLT Intrusions Delayed Recall	0.25	0.46	0.28	0.45	0.31	0.54	0.16	0.37	0.92	0.40

*Note: * indicates significant difference at $p \leq 0.05$.*

Appendix 9. CVLT-II Normative Data.

	Females (n = 57)		Males (n = 24)	
Measure	Mean	SD	Mean	SD
Trial 1 Correct	6.93	1.61	6.03	1.64
Trial 2 Correct	9.12	1.98	9.63	2.10
Trial 3 Correct	10.95	2.56	11.46	1.82
Trial 4 Correct	11.82	2.25	11.75	2.25
Trial 5 Correct	12.56	2.05	12.29	1.99
Total Learning Correct	51.39	8.24	51.33	7.83
Learning Curve	5.75	1.81	6.25	2.24
Trial B Correct	6.42	1.95	6.00	1.59
Immediate Free Recall	10.81	2.89	11.08	2.84
Immediate Cued Recall	11.47	2.44	12.08	2.08
Delayed Free Recall	11.18	2.64	11.38	2.10
Delayed Cued Recall	11.65	2.53	12.13	2.07
Recognition True Positives	14.37	1.62	14.88	1.26
Recognition False Positives	1.44	2.20	1.13	2.33
Semantic Clustering Ratio	0.246	0.133	0.241	0.108
Serial Clustering Ratio	0.235	0.103	0.231	0.108
Percent from Primacy	0.276	0.066	0.295	0.038
Percent from Middle	0.443	0.065	0.450	0.050
Percent from Recency	0.277	0.065	0.256	0.050
Total Learning Intrusions	1.33	2.06	0.71	0.86
Total Learning Repetitions	4.09	3.99	3.75	2.89

Appendix 10. HVLT-R Normative Data.

	Females (n = 57)		Males (n = 24)	
Measure	Mean	SD	Mean	SD
Trial 1 Correct	6.96	1.60	6.92	1.79
Trial 2 Correct	9.23	1.55	9.33	1.66
Trial 3 Correct	10.14	1.30	10.17	1.46
Total Learning Correct	26.33	3.58	26.42	4.24
Learning Curve	3.18	1.75	3.25	1.57
Delayed Free Recall	9.40	1.72	9.08	1.91
Recognition True Positives	11.18	1.71	11.29	1.12
Recognition False Positives	0.47	0.87	0.42	0.58
Semantic Clustering Ratio	0.340	0.182	0.332	0.173
Serial Clustering Ratio	0.238	0.137	0.235	0.168
Percent from Primacy	0.352	0.061	0.371	0.063
Percent from Middle	0.317	0.058	0.300	0.060
Percent from Recency	0.333	0.079	0.329	0.084
Total Learning Intrusions	0.49	0.98	0.42	0.72
Total Learning Repetitions	1.35	1.29	1.92	2.50

Appendix 11. RAVLT Normative Data.

	Females (n = 57)		Males (n = 24)	
Measure	Mean	SD	Mean	SD
Trial 1 Correct	6.60	1.53	7.29	1.57
Trial 2 Correct	9.51	2.05	9.54	2.13
Trial 3 Correct	10.96	1.99	11.08	2.36
Trial 4 Correct	11.84	1.99	11.71	2.18
Trial 5 Correct	12.60	1.82	12.38	2.37
Total Learning Correct	51.51	6.99	52.00	9.11
Learning Curve	6.00	2.28	5.08	2.45
Trial B Correct	6.78	1.68	6.33	1.81
Immediate Free Recall	11.04	2.62	10.46	3.18
Delayed Free Recall	10.79	2.94	10.75	3.37
Recognition True Positives	13.81	1.49	13.92	1.79
Recognition False Positives	1.31	1.80	1.25	1.85
Serial Clustering Ratio	0.351	0.154	0.396	0.133
Percent from Primacy	0.292	0.061	0.286	0.048
Percent from Middle	0.406	0.063	0.407	0.041
Percent from Recency	0.301	0.046	0.304	0.068
Total Learning Intrusions	0.60	1.32	1.08	1.50
Total Learning Repetitions	3.07	2.80	4.42	4.13