2019

An Analysis and Comparison of WMS-III and WMS-IV Verbal Paired Associates: Practical Implications for Neuropsychologists

Anthony Paul Andrews
Nova Southeastern University, cognosco33@yahoo.com

AN ANALYSIS AND COMPARISON OF WMS-III AND WMS-IV VERBAL PAIRED ASSOCIATES: PRACTICAL IMPLICATIONS FOR NEUROPSYCHOLOGISTS

by

Anthony Paul Andrews

A Dissertation Presented to the College of Psychology
of Nova Southeastern University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

NOVA SOUTHEASTERN UNIVERSITY

2019

Dissertation Approval Sheet

This dissertation was submitted by Anthony Paul Andrews under the direction of the Chairperson of the Dissertation committee listed below. It was submitted to the College of Psychology and approved in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Clinical Psychology at Nova Southeastern University.

9/16/19
Date of Defense

Approved:

John Klein Ph.D for Charles Golden
Charles Golden, Ph.D. Chairperson

Barry Nierenberg
Barry Nierenberg, Ph.D. Committee Member

Barry Schneider
Barry Schneider, Ph.D. Committee Member

9/16/19
Date of Final Approval

John Klein Ph.D for Charles Golden
Charles Golden, Ph.D. Chairperson

Acknowledgements

I would like to express my sincerest gratitude to my parents, whose love and constant support have been invaluable in my educational success. To Dr. Charles Golden, thank you for always leading by example and for your continued encouragement and commitment to neuropsychological excellence. To the committee members, Dr. Barry Nierenberg and Dr. Barry Schneider, thank you for your helpful feedback and ongoing support throughout the entire dissertation process.

TABLE OF CONTENTS

LIST OF TABLES

AN ANALYSIS AND COMPARISON OF WMS-III AND WMS-IV VERBAL PAIRED ASSOCIATES: PRACTICAL IMPLICATIONS FOR NEUROPSYCHOLOGISTS

by

Anthony Paul Andrews

Nova Southeastern University

ABSTRACT

The current study examined the performance of clinical outpatients on the Verbal Paired Associates (VPA) subtest from current and prior versions of the Wechsler Memory Scale (WMS). It was predicted that VPA from the WMS-III (VPA3) would be similar in agreement to the WMS-IV (VPA4) and that VPA4 would show a stronger relationship than VPA3 with intellectual abilities, sustained attention, and executive function abilities as assessed by the WAIS-IV; CPT-2; and the TMT, Category Test, and Stroop, respectively. Thirty-six adults were administered both the WMS-III and the WMS-IV, along with the other measures as part of a larger neuropsychological battery. Data were collected from an archival database of individuals clinically referred to an outpatient, university-based neuropsychology clinic.

Results of the current study showed that agreement for scaled scores was lower than expected for both VPA I and VPA II across versions of the WMS. Despite the relatively modest levels of strict agreement, current results did find that 89% of scaled score pairs fell within three scaled score points (i.e., one standard deviation) for both VPA3 and VPA4, and this was reflected in the magnitude of correlation coefficient, which was .76 in the predicted direction.

Results of the study also showed that VPA I and II across versions of the WMS-III and WMS-IV predict neuropsychological functioning similarly. Using the methodology described by Lee and Preacher (2013), direct comparisons found no significant differences between VPA3 or VPA4 in their ability to predict intellectual, attentional, or executive functioning abilities at the .01 level of prediction. The small sample size of the current study, the conservative alpha cut-off used, and the inherent cognitive heterogeneity inherent in a convenience sample of outpatient adults may have contributed to a lack of statistical power to detect real differences and masked otherwise significant differences between the tests.

Implications of the current study may be relevant for research and/or clinical applications. First, it is concerning that the degree of scaled score agreement is low across versions of VPA despite overall similar mean scores, and this low rate of agreement within each participant suggests that the tests themselves may be measuring memory in important, though poorly understood, ways. Overall, participants in the current study obtained identical scores across versions of the test less than half the time. One recommendation is to take caution when comparing results of VPA3 to VPA4 in serial assessments, such as with patients diagnosed with a neurodegenerative disease or litigants assessed by multiple practitioners to inform a trier of fact about a specific cognitive issue at stake in a matter being litigated. This will be become more relevant when WMS-5 is released in the next few years and most serial assessment will involve comparisons to VPA4. Current results also support the movement towards more transparent and thorough comparison of normative data for clinical populations, along with a specific investigation into the rationale behind future changes to VPA, namely the psychometric approach vs.

the theory-driven, neuroanatomically informed approach. For example, while adding semantically related, “easy” word pairs to VPA4 may have achieved a goal of normalizing mean scores on VPA4, it is unclear whether this psychometric victory represents meaningful changes in terms of how memory works in clinical populations, many of whom have been demonstrated to process information in a way dissimilar to non-clinical populations. The current research indicates the need for neuropsychologists to practice mindfully when using VPA to assess memory. Examples include adding alternative auditory memory measures to the test battery and conducting additional research with specific clinical populations to understand how performance on the test relates to in vivo memory functioning.

Statement of Original Work

I declare the following:

I have read the Code of Student Conduct and Academic Responsibility as described in the Student Handbook of Nova Southeastern University. This dissertation represents my original work, except where I have acknowledged the ideas, words, or material of other authors.

Where another author's ideas have been presented in this dissertation, I have acknowledged the author's ideas by citing them in the required style.

Where another author's words have been presented in this dissertation, I have acknowledged the author's words by using appropriate quotation devices and citations in the required style.

I have obtained permission from the author or publisher—in accordance with the required guidelines—to include any copyrighted material (e.g., tables, figures, survey instruments, large portions of text) in this dissertation manuscript.

Anthony Paul Andrews

September 16, 2019

CHAPTER 1
STATEMENT OF THE PROBLEM

Human memory is complex, vast topic. While advances in neuroimaging and other technologies have provided many insights into how our brains encode, store, and retrieve information, the neuroanatomical and theoretical specifics of how memory functions in humans are far from clear (Ekstrom & Ranganath, 2017; Jonides, Smith, Koeppe, Minoshima, & Mintun, 1993; Ranganath, 2010), and answers to basic questions about memory functioning are often estimates, and even those are often hotly debated. For example, no one really knows the storage capacity of long-term memory in humans. Recent research suggests it is at least 1 petabyte, about the amount of information stored in the entire internet, but others suggest the capacity may be much higher (Chaudhuri & Fete, 2016). Even our understanding short-term/working memory capacity, which was thought to be well-understood by the mid-1950s to be seven plus-or-minus-two (Miller, 1956), remains murky, for example, with recent research suggesting that capacity may be much higher and vary person-to-person depending on preexisting knowledge (Brady, Störmer, & Alvarez, 2016).

Neuropsychologists are tasked with assessing memory functioning in the context of this ever-evolving knowledge base. It is unsurprising then, that the tests and methodologies used to assess memory have changed drastically over the years. Wechsler's Memory Scales have been the most commonly administered memory tests since the introduction of the first WMS in 1945 (Wechsler, 1945). Assessment of memory performance is essential during the neuropsychological evaluation because of the fundamental role memory plays in cognition. Suspected deficits in episodic memory

prompt the most referrals to neuropsychologists, who assess memory performance to identify deficits, make differential diagnoses, and provide treatment recommendations. Episodic memory gives humans the ability to remember past happenings. It is the one component of memory that distinguishes humans from other animals, and it is vital for day-to-day functioning. It is a higher-level type of memory, in the sense it "evolved" from semantic memory, per Tulving, who coined the terms (2002).

The primary means of assessing memory performance is through standardized testing, and the most utilized memory battery is the Wechsler Memory Scales, now in its fourth edition (WMS-IV; Wechsler, 2002). In a recent survey of practicing neuropsychologists, Rabin, Paolillo, and Barr (2016) found that 62% selected the WMS-IV (or a prior version of the WMS) most often to assess memory. The dominance of the Wechsler scales is not a recent occurrence; in 2001, the WMS-III was chosen for memory evaluation by 71% of neuropsychologists (Rabin, Barr, & Burton, 2005; Wechsler, 1997). While changes to VPA from WMS-III to WMS-IV were well-intentioned responses by the publisher to the larger body of criticisms to WMS-III, and the test has been received well by practicing neuropsychologists (Rabin et al., 2016), the effect of these changes remains unclear.

CHAPTER II
REVIEW OF THE LITERATURE

David Wechsler's goal in developing the Weschler Memory Scale (WMS; Wechsler, 1945) was to create a test that would provide a "rapid, simple, and practical memory examination" (p. 87). In its original form, the WMS contained seven subtests assessing a variety of functions related to memory performance: Personal and Current Information, Orientation (purpose of both was to rule out aphasia and dementia), Mental Control (non-motoric processing speed measure), Logical Memory (immediate recall only), Memory Span (digit span forwards and backwards), Visual Reproduction, and Associate Learning. The last test, Associate Learning, is the oldest form of what is known on the current version of the WMS as Verbal Paired Associates. It consisted of 10 word pairs Wechsler first used in his study of patients diagnosed with Korsakoff's disease (Wechsler, 1917). Some of the paired associates were easy (i.e., semantically related), and some were hard (i.e., not semantically related).

Wechsler Memory Scale (WMS)

Standardization of the WMS was completed over 10 years at Bellevue Hospital using the results of "approximately 200 normal subjects, ages 25 to 50, both men and women" (p. 88). Age group means were calculated at 5-to-10-year intervals, and Wechsler provided a mathematical procedure for calculating the WMS Memory Quotient (MQ) by adjusting for performance on the WAIS (D. Wechsler, 1955). Wechsler noted that advantages of the WMS including its brevity (i.e., it took only 15 minutes to administer), standardization, age correction, and the ability to compare memory quotient to intelligence quotient. Wechsler concluded the WMS "should be useful in detecting

special memory defects in individuals with specific organic brain injuries and may prove of concrete value in the examination of some of the soldiers and sailors returning with head injuries" (p.90).

Early reviews of the WMS were general, brief, and for the most part, lacked statistically sophistication. Nevertheless, the earliest research examining WMS performance within and across groups revealed numerous potential shortcomings. For example, (Cohen, 1950) found that the test was unable to meaningfully distinguish between patients with organic brain damage versus those with schizophrenia or those diagnosed as psychoneurotic (Howard, 1954) .Howard conducted a more thorough study examining the test ability to distinguish between patients with organic brain damage versus those who were "mentally disturbed" but had not been diagnosed with organic brain damage (p. 377). Participants in the experimental group were diagnosed with encephalitis, epilepsy, or paresis, and participants in the control group were diagnosed with a psychotic disorder. Results of the study showed that the WMS failed to differentiate between encephalitis and epileptic groups and their control groups. However, the WMS did differentiate between paretics and matched controls "above the 1 per cent level of confidence [on] Memory Quotient, Personal and Current Information, Orientation, Total Mental Control, Counting by Threes, and Visual Reproduction" (p. 380). Howard concluded that severe brain damage was required before the WMS would be sensitive enough to differentiate between organic and psychogenic patients. He did not, however, seem to consider the role that psychosis plays on attention (and therefore memory), and so his results may have been confounded.

In 1966, Howard found 31 of his patients from the 1950 study published a 15-year

follow-up examining memory performance as assessed by the WMS of organics versus non-organics. Howard's metric of change, as with his previous study was the Memory Quotient (MQ). His results show that after 15 years, that while 14 of the 19 experimental patients showed decline in memory functioning (median MQ drop of six points), the results were not statistically significant overall. Of all subtests, only one item Visual Reproduction decreased at the 5% level of significance (note that this was one item on the test, which measured the design after a 10 second observation period). Howard noted that the lack of change in results could have been because of age corrections for MQ. Regarding nonorganic, Howard found that the results were similar to the experimental group and were nonsignificant other than two items on Visual Reproduction. Overall, he found that the groups differed significantly on more subtests in 1950 (i.e., MQ, Personal and Current Information, Visual Reproduction Total, C-1, and C-2) than in 1966 (i.e., Personal and Current Information and C-2). Howard concluded that the similarities in memory performance over time and continued decline in memory functioning was a result of long-term hospitalization rather than organic versus nonorganic factors (Howard, 1966).

In 1958, a study examining diagnostic and predictive accuracy of the WMS in older adult psychiatric patients was published (Walton, 1958). The study followed a group of geriatric patients who had been administered the WMS on four separate occasions to track the efficacy of a vitamin intervention. For the study, they were readministered the WMS two years later. Unlike Howard's research referenced above, Walton referenced in accounted for environmental effects on memory test performance in his 50 patients. He noted that "the initial diagnosis and the first [Memory Quotient]

assessments were clearly not very reliable, though the results of the fourth assessment corresponded most closely with the final diagnosis" (p. 1113). Walton noted that the length of time hospitalized probably led to an apathetic state in many of the patients, which may have affected their performance. After a careful review of their records, performance, and final diagnosis, Walton concluded that patients accurately diagnosed as organic showed little improvement over the successive re-testings. On the other hand, those accurately diagnosed as having an affective disorder showed "large differences between first and final [Memory Quotients]" (p. 1113). Walton noted that depression and apathy caused by lengthy hospitalizations could cause misdiagnosis of dementia due to poor WMS performance, but that accurate diagnosis of patients revealed distinct WMS memory profiles between the organic and nonorganic groups.

Statistically sophisticated analysis of the WMS did not begin until the early 1970s when the first published factor analysis of the WMS was published in 1970, 25 years after the test had been in widespread use. The term sophisticated is relative, for the entire analysis was less than one page. Data for the factor analysis were collected from 622 patients at the Mayo Clinic over the course of seven years and ranged in age from 15 to 87 years (Davis Jr. & Swenson, 1970). Initial results revealed two factors, and oblique rotation revealed a factor pattern involving a first factor that included long- and short-term storage and retrieval, which the authors named Memory. Test loading on the first factor included Information, Orientation, Logical Memory, Visual Reproduction, and Associate Learning. The second factor included what the authors described as a "freedom from distractibility" factor (p.430) and included the subtests Mental Control, Digits Forward, and Digits Backward.

In 1971, Dujovne and Levy acknowledged the unfavorable results of prior attempts at research to validate the WMS, and they suggested that ignorance of its underlying psychometric properties could be a contributing factor: “On the whole, the results of validation studies have been confusing mostly and negative, perhaps due to the fact that investigators proceed to validate the effectiveness of the test as a diagnostic tool without knowing its structure” (p. 351, Dujovne & Bernard, 1971). Theirs was the first published study to systematically investigate the underlying structure of the WMS, and thus it will be discussed here in some detail. The authors included a robust, highly functioning, normal sample (276 nonclinical persons, aged 16 to 71) with mean WAIS FSIQ scores of 115 and mean WMS MQ scores of 115. Slight differences in scores were noted for males and females (e.g., MQ SS = 115.8 and 114.8, respectively). Experimental group data were collected from three clinics and included a mostly male population (85% male) consisting of mostly acute and chronic brain disorders (N = 81), psychotic disorders (N = 21) psychoneurotic disorders (N = 29), and some form of personality disorder (N = 35). All participants in the study had completed the WMS and the WAIS.

Results showed items on the Personal and Current Information and Orientation subtests were passed by both groups, so those items were eliminated from subsequent analyses. A factor analysis using the verbal IQ and performance IQ scores as reference variables revealed three factors for the normal sample. Factor I accounted for 51% of the common variance and was named general Retentiveness, and was comprised of Digits Backward (.68), Count Backward (.64), Digits Forward (.64), Say the Alphabet (.61), Logical Memory B (.57), and Visual Reproduction C-1 (.44). Given its strong loadings with verbal (.85) and nonverbal (.79) intellectual abilities, the authors concluded that the

factor "may be tapping (a) general intelligence rather than memory, or (b) general retentiveness or general memory functioning important for successful intelligence test performance" (p.352).

Factor II was named Simple Learning, accounted for 27% of the common variance, and was made up largely of easy items from Associate Learning, including: Second Easy Associations (-.88), Third Easy Associations (-.88), and First Easy Associations (-.67. Simple learning was not strongly related to intelligence and per the authors, "is composed of the associations laid down a childhood and reinforced throughout life" (p.352).

Factor III, named associative flexibility, accounted for 22% of the common variance and was comprised of hard items from Associate Learning. Items included Second Hard Associations (.86), First Hard Associations (.63), and Third Hard Associations (.63). Per the authors, "the task involved in the hard associations requires the formation of entirely new and unfamiliar associations" (p. 352). They also noted that the only reference variable that loaded onto Factor III was the Digit Symbol subtest (.39), and they noted, "In spite of the fact that the Digit Symbol deals with visual stimuli and the Hard Associations with verbal stimuli, a common feature is shared by the two subtests inasmuch as they deal with pairs of variables that do not stand in logical or meaningful connection" (p. 352). They also acknowledged the influence of age on both digit symbol and the hard items of Associate Learning, and they describe the results as "not surprising […] Since the Digit Symbol subtest is known to decline earlier and to drop off more rapidly with age than other subtests of intelligence. The same may apply to the Hard Association subtest" (p. 352). Anecdotally, this insight, described nearly 50

years ago, is interesting given the changes to subsequent versions of WMS, specifically WMS-III, when the decision was made to drop the easy items, which resulted in the introduction of unintentional floor effects to the Verbal Paired Associates subtest.

The primary factor analysis by Dujovne and Bernard of the patient sample also revealed three factors. Factor I accounted for 31% of the common variance and was named Mental Control. The authors noted that it was more related to verbal intelligence (.64) than to performance intelligence (.55) and it included Count by 3's (.64), Say the Alphabet (.63), Digits Backward (.58), Digits Forward (.51) and Count Backward (.51). It is noteworthy that reference tests Arithmetic and Digit Span loaded highly onto Factor I. Unlike Factor I, General Retentiveness, from the normal sample, however, Factor I from the patient sample was not equated with overall intelligence. The authors noted, however, "Mental Control requires a relative autonomy of the subjects intellectual functioning from the disturbing effects of anxiety and/or psychic impairment due to neurological dysfunction [and therefore] it would be expected to play an important role in intelligence test performance."

Factor II from the patient sample accounted for 32% of the common variance and was named Associative Flexibility. This factor loaded strongly on Associate Learning subtests: Second Easy Associations (-.85), Third Easy Associations (-.71), First Easy Associations (-.68), Third Hard Associations (-.63), Second Hard Associations (-.63), and First Hard Associations (-.40). Like Factor I from the patient sample, associative flexibility variables were not related to intelligence. Regarding the combination of easy and hard items on this factor the authors suggested "unlike the normal sample, the presence of the hard and easy associations on the same factor in the patient group would

suggest that a certain amount of the associational flexibility involved in the hard associations performance is needed to establish connections of the easy type in the patient group" (p. 353).

Factor three from the patient group accounted for 37% of the common variance, was named cognitive dysfunction, and included the following variables: Logical Memory B (-.74), Logical Memory A (-.66), Visual Reproduction B (-.56), Visual Reproduction C-1 (-.53), Visual Reproduction C-2 (-.51), Third Hard Association (-.48), First Hard Association (-.47), and Second Hard Association (-.46). The authors noted that Factor III variables required meaningful processing of "complex" material. Regarding memory function, the authors suggested that Factor III variable dealt more with the retention of information rather than learning of information, and they used this rationale to explain the loadings of hard items on both Factors 2 (i.e., learning) and 3 (i.e., retention). The authors also noted that items on Factor were the most for the patient group, perhaps due to impaired ability to synthesize information/see the gestalt, as evidenced by the fact that all the loadings (other than Hard Associations) "have a form." The reasoned that the presence of the Hard Associations on Factor III was evidence that the participants in the patient group were handling the "parts of the gestalt as discrete units" (p. 353). They also suggested that unlike the normal group, participants in the patient group approached items on Hard Associations and Logical Memory in a concrete manner. The authors concluded by suggested that items on the WMS be regrouped into 3 subscales reflecting their factorial loadings, and they also suggested that scores on the three resulting subtests should be substituted for MQ. One shortcoming of this study was the difference in intellectual ability between the two groups. The normal group was approximately one

standard deviation above the mean for intellectual and memory performance, which could lead to a lack of generalizability. This may lend more weight to the results from the clinical sample, though performance here was confounded by the presence of organic brain dysfunction and mental illness.

In 1975, six parallel forms for the Associate Learning subtest were introduced (Nott, 1975). Eight subjects with a range of organic and mental health diagnoses were used to validate the tests. Each subtest was patterned after the WMS subtest, including the location within the subtest of easy and difficult items. The purpose of the parallel forms was presumably to allow for assessment of progress over time.

In 1973, Kear-Colwell published a study investigating the structure of the WMS and its relation to intelligence and various CNS lesions. Their sample included 250 patients referred from a general hospital for referral in a clinical psychology department to assess cognitive functioning. All participants in the study had included the WMS Form I (Wechsler, 1945) and the WAIS (Wechsler, 1955). Mean age was 47.76 (SD = 12.73) and the sample was 65% male. 66 patients had confirmed organic disorder (i.e., head injury, dementia, Parkinson's disease, CNS tumor, epilepsy, etc.), while 184 did not (though this information was not always known prior to the assessment). Mean IQ scores were 95.04 (SD = 14.65), and mean MQ scores were 98.39 (SD = 19.53).

The sample was divided into two groups (i.e., Confirmed Organic and Not Confirmed Organic), and the results of a principal factor analysis revealed factors "of almost identical structure" (p. 386). As a result, all participants were included in a factor analysis. Factor I included high loadings on Logical Memory, Visual Reproduction, and Associate Learning. Factor II loaded on Mental Control, Digit Span, and Information,

while Factor III loaded on Orientation and Information. The authors next obtained sten scores for each factor and factor analyzed these along with WAIS scores, age, and sex. The results revealed three factors, which accounted for 74% of the common variance. Factor A seemed to measure intellectual ability and included high loadings on the three WMS factors along with FSIQ, Verbal IQ, and Performance IQ. Factor B was described as "Verbal-Performance Discrepancy" and was weakly related to any of the WMS factors (r range = -0.05-0.20). Factor C was age, and the high correlation with WMS Factor I was interpreted by the author as "a tendency for the ability to learn and recall new material to deteriorate with age" (p. 387). Citing the results of the factor analyses, the author reported that the structure of memory is the same for persons with a without brain dysfunction, regardless of intellectual ability or age. Like (Dujovne & Bernard, 1971), Kear-Colwell suggested interpreting WMS results using three factors rather than the overall MQ (Kear-Colwell, 1973).

In 1977, Kear-Colwell conducted a replication study on 112 patients to replicate the 1973 results (Kear-Colwell, 1977). The 1977 sample was independent from the 1973 sample, though the patient characteristics, including referral information, were similar, with exception of age and organicity (higher percentage of confirmed organic pathology in the 1977 sample). Results revealed three WMS factors which accounted for 81% of the common variance. A high correlation was observed between actual factors scores from the replication sample and those obtained in the prior study, which the author commented on: "These correlations of factor scores are the crucial test of factor congruent in any cross validation of factor structure and meaning […] These findings indicate high reliability for the three factors and by implication suggest high validity" (p. 484). Kear-

Colwell then combined the sample from the replication and original study (N = 362), repeated the factor analysis procedure, and again found the same three factors, which accounted for 76% of the common variance. The author noted that Visual Reproduction correlated highly in both studies with Verbal IQ from the WAIS (.59 and .48, respectively), and noted that the task may be verbally medicated. The author also found that separating Digit Forwards from Digit Backwards "add very little unique information to the factor structure," and "A single Digit Span score was adequate" (p. 485).

In 1978, Kear-Colwell and Heller attempted to replicate the results of the prior factor analyses on a nonclinical sample and to examine the effects of age, sex, and social class on memory performance (Kear-Colwell & Heller, 1978). The sample included 116 Health Service employees. Samples were stratified by age, sex, and social class (assessed by occupation). Results of the factor analysis revealed support for Factors I and II from the prior study. Factor II was not fully supported, as it was broken into two subfactors, likely because of discrepancies in performance on Information and Orientation from the prior combined patient sample and the current nonclinical sample. However, the overall common variance accounted for by the 1978 results was very similar to those obtained previously. When younger adults (below 35 years) and older adults (35 years and above) were compared, significant age effects were found for Logical Memory, Visual Reproduction, Associate Learning, Total Raw Score, and Factor I at $p < .001$. Differences were found for Digit Span at $p < .01$. Differences were found for social class effects for Factor I, Factor II, MQ, and Total Raw Score at $p < .001$. Differences were found for Factor III at $p < .01$. Males and females performed similarly on Factors I and III, but males performed better on Factor II ($t = 2.2$, $p < .05$), specifically on the Digit Span

subtest ($t = 2.8$, $p < .01$). No second differences were found for Mental Control.

Following the increased interest in researching the WMS, Prigatano published a literature review of the WMS paying attention to its psychometric qualities and clinical utility (Prigatano, 1978). In his review, Prigatano noted, "Despite its widespread use, many do not consider it a good psychometric instrument (p. 817). He went on to describe several different problems (most of which have been described earlier), including:

"(1) a relatively small and restricted standardization sample and consequently inadequate norms (Wechsler, 1945); (2) little information about the reliability of the test (Stone & Wechsler, undated); (3) disagreement over the factor structure of the WMS (Davis Jr. & Swenson, 1970; Dujovne & Bernard, 1971; Kear-Colwell, 1973) and consequently its validity; and (4) the meaning of the Memory Quotient and whether it measures something other than IQ (e.g., Fields, 1971)" (p. 817).

Prigatano (1978) noted that Logical Memory scoring was difficult because no specific instructions were provided as to what constitutes a "correct idea" (p. 818). He suggested revising the scoring of Logical Memory to include specific information recalled and that account for guessing. Regarding the norms, Prigatano noted that they had not been updated since Wechsler's original 1945 sample.

Regarding the psychometric properties of the WMS, Prigatano noted several problems with its reliability. First, he noted that no information had been published regarding the data Weschler had collected over the 10 years he was developing the WMS. He also noted that no alternate-form or test-retest reliability estimates had been reported by Wechsler, and that those had followed had been inadequate for the most part. He did

find anecdotal support for test-retest reliability of the WMS, primarily based on the research by Howard described earlier (Howard, 1966).

Regarding the factor structure of the WMS, Prigatano referenced those described above plus two others and noted "there are at least two and possibly three factors that compose the WMS when it is given to a combined group of neurologically and psychiatrically impaired patients." He then went on to describe the factors referenced earlier.

Regarding the utility of the WMS, Prigatano (1978) noted that the test had limited validity in its current form and that the primary value was obtaining a MQ, which could then be compared to a patient's FSIQ. For patients with average or below-average intelligence as measured by FSIQ, a difference in MQ of 12 or more points was enough to diagnose a short-term verbal memory deficit. He added that "in each individual case, however, other supporting data would be needed to confirm this suspicion." Prigatano (1978) concluded "the WMS needs to be improved substantially […] The best thing that can be said […] is that [the WMS] often reflects amnestic disturbances associated with left cerebral hemisphere lesions." He also added that it "appears to be generally insensitive to memory disturbances associated with the nondominant (right) cerebral hemisphere."

In 1981, the revised edition of the WAIS was released (WAIS-R; Wechsler, 1981), and researchers quickly noted that comparisons between WAIS-R FSIQ and MQ were not the same as WAIS FISQ and MQ. Prifitera and Barley (1985) examined WAIS versus WAIS-R performance in comparison to WMS MQ scores in 120 psychiatric inpatients and found that mean WAIS FSIQ was higher (M = 102.93, SD = 16.29) than

mean WAIS-R FSIQ ($M = 96.15$, $SD = 13.30$, $t(118) = 2.50$, $p < .02$. The authors concluded the finding "probably reflects a difference between the norms of the two intelligence scales rather than a true individual difference between groups" (p. 565). They also found that unlike WAIS FSIQ, WAIS-R FSIQ was likely to be lower than MQ, and they concluded, "the 12 point rule of thumb is not applicable and comparing WAIS-IV FSIQ with WMS MQ" and they urged that the WMS be renormed because "MQs are inflated compared with the WAIS-R FSIQ" (p. 565).

Throughout its history, numerous versions of the WMS were used by clinicians to assess memory performance. As a criticism of the WMS grew, modifications to the test became commonplace to account for perceived shortcomings. It is noteworthy that considering the numerous valid criticisms described by Prigatano (1978), he failed to note one of the most important short-comings of the WMS: the lack of attention to delayed recall. Though it was explicitly addressed in later editions of the WMS, the most popular modification to the WMS was made by Russell (1975), who cited the research described earlier by Dujovne and Levy (1971) as a major shortcoming of the test (namely, that the WMS assume memory is a unitary construct). Russell, in referencing Luria (1973), also noted that the test implicitly assumes that the brain functions as a whole rather than as a concert of localized processes. Russell also described the research of Milner (1968), who demonstrated that memory differences exist between cerebral hemispheres.

In referencing the research of Kesner (1973), Russell described a need for assessment of long-term memory functioning in addition to short-term memory funcitoning. Russell's goal was to create modifaction to the scoring system of the WMS,

taking into consideration the above-referenced research, to create "a memory scoring method that will measure more precise types of memory and still be a clinically useful tool" (p. 800). He relied on the factor analysis studies by Kear-Colwell, which were described in detail earlier, and identified Factor III as containing the variables best able to differentiate between mild brain damage from a group of non-brain-damaged patients. As the reader may recall, Factor III consisted of Information, Logical Memory, Visual Reproduction, and Associate Learning. For various reasons, Information and Associate Learning were removed from his scoring system (e.g., they did not add much additional information), and only Logical Memory and Visual Reproduction were retained in the scoring system.

Russell also advocated for a long-term memory measure, which he accomplished by re-administering the memory test 30 minutes after the first administration. Finally, Russell created a way to measure the amount of memory impairment by coordinating memory impairment scores with the Average Impairment score on the Halstead-Reitan battery (Russell, Neuringer, & Goldstein, 1970). With this new scoring method, "patterns of memory impairment produced by brain damage can be derived. Thus, this new method of scoring includes measures of the relative amount of lateralized impairment for both short- and long-term memory related to the Halstead-Reitan battery" (p. 803).

In 1986, Ernst and colleagues investigated a version of the WMS that included 30-minute delayed recall scores (percent retained) for Logical Memory and Visual Reproduction. Also, unlike Russell's system, which they referenced, Ernst and colleagues included Associate Learning because "our clinical observations […] suggests that Associate Learning taps a different type of verbal recall than Logical Memory […] a

delayed recall of Associate Learning has been clinically useful to compare patients' retention on the two different types of verbal learning abilities" (p. 310). In describing the differences between Logical Memory and Associate Learning, the authors stated, the "hard" (non-associated) and "easy" (highly associated) items […] offer two levels of associational cues to ease encoding and prompt retrieval. The WMS Logical Memory stories provide an even richer verbal context to facilitate encoding of information" (p. 310).

Participants in Ernst and colleagues' study included 70 adults referred for psychological testing ranging in age from 18 to 66 (M = 34.3, SD = 12.3). Principal factor analysis followed by orthogonal rotation using the varimax criterion was conducted. Four factors were identified: Factor I included subtests related to attention and learning/recall (i.e., Mental Control, Digit Span, Visual Reproduction, Logical Memory and corresponding delay); Factor II included immediate and percent retained scores for Selective Reminding and both "easy" and "hard" items from Associate Learning, and the authors described this factor as measuring repetition learning. Factors III and IV were minor, with Eigenvalues less than 1.0. However, the authors noted, "the striking findings here are the loadings of visual reproduction and corresponding percent retained on separate factors (1 and 4, respectively)." The authors suggested than this finding supported earlier research by Larrabee, Kane, Schunck, and Francis (1985) showing that delayed recall is superior to immediate recall when assessing memory using Visual Reproduction.

Regarding the findings that Mental Control, Digit Span, Logical Memory, and Visual Reproduction loaded highly on the same factor, the authors suggested that those

tasks do not require repetition and rely more on attention and concentration than Associate Learning. The authors also noted that the factor loadings may have treatment implications, with better performance on Associate Learning than Logical Memory indicating the use of "repetition, feedback, and prompts in working with the patient" (p. 313). In contrast, the opposite finding would indicate that "information and recommendations offered to patient would likely be effectively recalled even with single presentations if the material relevant and understandable to the patient" (p. 313).

Treatment implications for better visual than verbal recall after a delay were straightforward: use visual associations when working with the patient. The authors concluded by noting that their results provided support for the decision to include delayed recall of Associate Learning in the battery. They noted that both "hard" and "easy" items loaded on the same factor, suggesting that Associate Learning "is a clear measure of rote learning and memory" and not hampered with the attentional demands associated with Logical Memory (p. 313).

Interim Summary

The WMS was amongst the first memory tests, and it stood the test of time. For over 40 years, it was arguably the most widely used memory test by neuropsychologists. Its similarity to the WAIS and WAIS-R, which were undoubtedly the most widely used instruments for measuring intelligence, probably accounted for its popularity. The test was subjected to criticism shortly after its publication and throughout its existence by researchers and clinicians for its content, (lack of appropriate) standardization, questionable reliability, and ecological and construct validity. Looking back, it is easy to cite numerous shortcomings of the WMS, but it also worth mentioning that our

understanding of memory, both anatomically and theoretically, was lacking when Wechsler published the WMS in 1945. In fact, the WMS was not based on any explicit theory of memory at all. Rather, it was simply designed to provide a "rapid, simple, and practical memory examination" (Wechsler, 1945, p. 87). The research reviewed thus far demonstrated that the WMS measured primarily verbal episodic memory and working memory/attention. In its original form, it only measured immediate recall, but later versions assessed long-term retrieval. By the time the WMS-R was released in 1987, researchers and clinicians had learning much more about the neuroanatomical underpinnings of memory, and the theoretical foundations of memory were better understood.

Wechsler Memory Scale-Revised (WMS-R)

The WMS-R (Wechsler) was released in 1987 and included "major revisions" that were completed prior to Wechsler's death in 1981. As described by Powel (1988), per the test authors, "extensive changes" were made from the WMS to the WMS-R, including: adult norms stratified at nine age levels; replacing of a single memory score (i.e., MQ) with five index scores (General Memory, Verbal Memory, Visual Memory, Attention/Concentration, and Delayed Recall); the addition of subtests assessing figural and spatial recall (i.e., Figural Memory, Visual Paired Associates, and Visual Memory Span); and improved scoring criteria. Overall, the WMS-R included 8 subtests plus a screening/mental status exam subtest. Four of the eight subtests assess recall after a 30-minute delay. Like the WMS, the composite scores that are derived from the eight subtests have a mean of 100 and a standard deviation of 15.

The structure and scoring of the WMS-R is more complicated than the WMS. The

General Memory score is based on the weighted sum of the Verbal Memory I and Visual Memory I, and subtest weights vary from one-to-two for each composite score. Numerous subtest changes were made from WMS to WMS-R, as described by Powel (1988). Information and Orientation subtest items tend to be answered correctly by most examinees, so its score is not included in the memory indices. Mental Control was unchanged in format, but speed credits were eliminated from the scoring system. Figural Memory, a new subtest introduced in WMS-R, was designed to measure recognition of abstract visual patterns. Powell (1988) noted that it could be useful in differentiating cortical from subcortical dementias, a reference to the idea described by Butters (1987) where patients with subcortical dementia would have intact recognition while those with cortical dementia would not. Powell (1988) noted several potential shortcomings of the new subtest, including a lack of scaled scores, the inherent difficulty of the test for normals making it difficult for elderly examinees, and no verbal analogue.

Logical Memory stories were “made equivalent in difficulty and equivalent in score obtainable” (p. 398). The scoring criteria were also improved, and a delayed recall condition was added. Visual Paired Associates, a new subtest, was added as a Visual Analog to Verbal Paired Associates Subtest. Six colors are paired with six nonsense drawings for at least three trials. The examinee is required to identify all color/drawing pairs to achieve “criterion,” though the task is discontinued after six trials even if the criterion is not reached. The score is derived from the first three trials, and a delayed condition is included.

Verbal Paired Associates was the revised name for Associate Learning from the WMS. Changes included the deletion of the two easiest word pairs, which left four easy

and four hard word pairs. Trials are continued until all pairs are learned in the identical manner to Visual Paired Associates. A delayed recall condition was also added. Changes to Visual Reproduction included modified item content, changes to the scoring system, and the addition of a delayed recall condition. Digit Span introduced easier items to both digits forward and digits backwards. As Powel observed, the actual numbers on WMS-R Digit Span are different from those found in the WMS and in the WAIS-R, which required examinees taking both the WAIS-R and the WMS-R to complete two different digit span subtests to calculate standardized scores. Visual Memory Span was a newly introduced subtest designed as a spatial analog to the digit span subtest.

There were several important changes to administration and scoring criteria from WMS to WMS-R. First, the administration time increase dramatically from about 15 minutes with Wechsler's version of the WMS to 45-60 minutes with WMS-R. Powel noted that changes to scoring criteria for Logical Memory and Visual Reproduction were "outstanding" with examples of 0- and 1-point responses, along with the removal of 1/2 point responses. This was reflected in excellent results for interscorer reliability on Logical Memory (0.99) and Visual Reproduction (0.97).

While the addition of the Delay Recall measure was welcomed, it was not without its own problems, as noted by Powel (1987). First, there was no breakdown between the verbal and visual components of the Delayed Recall index, the discovery that was probably disappointing for neuropsychologists who had been using Russel's' (1975) version of the WMS. Powell also noted that the Delayed Recall index includes no measure of information decay and as a result, "the delayed recall measure can be misleading since it does not take into account the patient's initial level of performance"

(p. 401). A crude work-around was suggested, though no normative data were available for comparisons.

Another shortcoming of the WMS-R was that some subtests had few items (e.g., Mental Control had only three), which contributed to the lack of subtest scaled scores available in the WAIS-R. Further, while means and standard deviations are provided for all subtests, Powel cautioned "their use is questionable […] based on my initial experience with these tasks, the underlined distributions may be highly skewed" (p. 401). He concludes that the combination of these factors makes interpretation of individual subtests difficult.

Regarding development of the WMS-R, Powel noted that the publishing company did a "professional job […] one nearly at the level of the WAIS-R" (p. 401). This is a noticeable improvement from the "approximately 200" patients Wechsler obtained with the original version of the WMS. The standardization sample was stratified by age and included approximately 50 case each for most of the nine groups. For the first time, demographic information was published for the sample, including sex, age, and region. To control for intellectual ability, to age bands were administered full WAIS-R's and the others were administered abbreviated versions of the WAIS-R. Another noticeable improvement from the WMS to the WMS-R was the inclusion of statistical analyses to establish relatability and validity of the WMS-R. As noted by Powel, "The reliabilities by age ranged from .27 (Figural Memory; age 55-64) to .93 (Attention/Concentration; age 35-44) with the majority of the composite reliabilities in the high 70's to low 90's" (p.402). Overall, Powel concluded the WMS-R "to be a significant improvement over its predecessor" for the reasons cited above (p. 402).

Studies of the factor structure of the WMS-R were included in the test manual and revealed a two-factor structure for normal subjects in the standardization sample: a general memory factor and a working memory factor (Wechsler (1987) as cited in R. A. Bornstein and Chelune (1989). However, when WAIS-R FSIQ was included in the analysis, the factor loadings changed, with the first factor containing the loadings for working memory and FSIQ. Shortly thereafter, Bornstein and Chelune conducted a series of factor analyses with normal and clinical samples to investigate the factor structure of the WMS-R (1988; 1989). In their 1988 study with 434 clinical patients referred from hospital settings, the authors found a two-factor structure for the WMS-R immediate memory subtests in isolation, like Wechsler (1987). However, when the WAIS-R was added, Bornstein and Chelune found a unique factor solution where most WAIS-R scores loaded on one factor, while scores from the WMS-R loaded onto a separate factor. When delayed recall subtests were added to the analysis, a three-factor solution was identified: 1) Verbal Memory, Nonverbal Memory, and Attention/IQ. The second study examined the same sample grouped by age (≤ 39 years; 40-55 years; and ≥ 56 years) and education (<12 years; 12 years; and > 12 years). Noteworthy findings included that most nonverbal memory tests, especially Visual Reproduction, loaded onto a nonverbal memory factor in the youngest group but on a verbal memory factor in the two oldest groups. The authors interpreted the finding as being consistent with earlier research demonstrating changes in memory associated with aging (e.g., Haaland, Linn, Hunt, & Goodwin, 1983), though they noted that the Figural Memory subtest did not follow this trend, possibly because it is a recognition memory test. Three factors were identified in each educational group when FSIQ was included in the analysis: 1) Verbal Memory, 2) Nonverbal Memory, and

3) IQ/Attention. The authors also found that the highest education group (i.e., > 12 years) had a different factor loading: IQ loaded onto the second factor, suggesting a link between intelligence and education level. At the lower levels, IQ and education did not predict WMS-R performance.

In contrast with the WMS, which enjoyed a "honeymoon" period relatively free from criticism for several decades until the 1970s, the WMS-R was attacked repeatedly shortly after its release. In 1989, Loring, Lee, Martin, and Meador conducted a study to determine the WMS-R's ability to predict laterality in patients who had underwent either right or left lobectomy. Although the sample size was small (most lobectomy studies are), Loring's study provided damaging evidence that the WMS-R was unable to adequately predict laterality using discrepancy scores between verbal and visual memory indices, even within the same group (i.e., with each participant serving as his or her own control). Loring concluded, "the WMS-R Verbal and Visual Memory Indexes should not be treated as equivalent to the brain structures whose functions they are designed to assess" (p. 201).

Loring (1989), in another critical review of the WMS-III, noted that the two new subtests – Figural Memory and Visual Paired Associated – were included with little rationale and that the tests "lack the necessary face validity to assess visual memory and learning" (p. 63). He also noted that Visual Paired Associates is confounded by "a significant verbal component," and that "almost all patients spontaneously employ verbal labeling" while taking the test" (p. 63). He also pointed to prior research demonstrating that visual paired associate tasks (even those that are not easily verbalized) do not lateralize to the right hemisphere and he suggested that other tests would have been better

suited, such as a facial recognition task. Loring was also critical of the Delayed Recall Index, noting that it did not included a delayed component for the Figural Memory subtest and that the subtests are weighted differently for the Delayed Memory and General Memory Indexes. Other criticisms included an unacceptably small and geographically restricted sample size and no normative group for people who did not graduate high school. He concluded, "The WMS-R still appears to be more a test of verbal learning […] It is unfortunate […] that the advancements made over the past several decades in cognitive and experimental/clinical psychology were largely ignored" (p.67).

In 1991, Elwood examined the WMS-R's psychometric properties concluded the standardization is "inadequate by current standards" due to a small sample size and sampling errors, both of which lead to increased standard error in the subtests and indices (p. 196). In reflecting on the reliability of the test, he noted that only the General Memory and Attention/Concentration indices "met even the most liberal standards for reliability," and only 4 of the 12 subtests were reliable (i.e., Digit Span, Visual Memory, and Logical Memory I and II)." Regarding factor analytic studies, his criticism was equally sharp. He noted that for clinical samples, only a General Memory factor was supported, as the second factor (described above) was conflated with IQ and so could not actually measure attention and concentration. He concluded that the WMS-R was a unidimensional test and that the multidimensional index scores were ineffective. He recommended that clinicians keep in mind the large standard errors of measurement for both the subtests and the indices, and that index scores should be reported as confidence intervals. He also recommended that the subtests not be interpreted in isolation, except for the four

mentioned earlier that were found to have acceptable reliability statistics.

In 1993, Burton, Mittenberg, and Burton conducted a study that was more supportive of the WMS-R, or at least its proposed multifactorial structure. They stated that the results found by Bornstein and Chelune (1988) were the product of exploratory factor analysis, and they argued that confirmatory factor analysis was required to makes causal determinations about factor solutions. They performed a confirmatory factor analysis using the WMS-R standardization data and tested numerous hypothetical factor models, including a one factor model of general memory, a two-factor model, such as proposed in the WMS-R manual (Wechsler, 1987), and the three-factor model proposed by Bornstein and Chelune (1988), amongst others. They also included a three-factor model based on the research of Roh, Conboy, Reeder, and Boll (1990), who found a three-factor model in a sample of head-injury cases consisting of Attention/Concentration, General Memory, and Delayed Recall factors. Roh and colleagues' contribution was methodological; by correlating the measurement error for immediate and delayed subtests, they effectively removed the method variance shared by the subtest conditions. Burton, Mittenberg, and Burton's study (1993) examined whether the three-factor solution would generalize to normal individuals (i.e., the WMS-R standardization sample). They tested the hypothesis using confirmatory factor analysis to evaluate the goodness of fit of seven commonly proposed solutions, including (a) a one factor solution (i.e., General Memory), (b) a two factor solution, such as was described by Wechsler (1987) in the WMS-R manual (i.e., Attention/Concentration and General Memory), the three factor model suggested by Robert A. Bornstein and Chelune (1988) (i.e., Attention/Concentration, Verbal Memory, and Visual Memory), (c) the model

implied by the WMS-R indices (i.e., Attention/Concentration, Immediate Verbal Memory, Immediate Visual Memory, and Delayed Recall), and several models by Roh and colleagues, including the three solution of Attention/Concentration, Immediate Memory, and Delayed Recall (1990). As predicted, the results of the maximum likelihood confirmatory factor analysis demonstrated that the three-factor solution described by Roh and colleagues best fit he WMS-R standardization data. The authors noted that the two-factor model in the WMS-R manual was hampered by the omission of the delayed recall subtests. The authors also used the results of Loring and colleagues (1989) to support their finding that no distinct verbal and visual memory factors existed on WMS-R (or at least their inclusion did not improve model fit).

In 1995, Gass described the importance of differentiating between memory storage and retrieval and published a multiple-choice recognition test for Logical Memory, along with a cueing technique for Visual Reproduction. He administered the new tests to 94 psychiatric inpatients and 99 brain-injured patients at the Miami V.A. Medical Center and found that the brain-injured sample performed worse than the inpatient sample on both subtests and that performance on the Visual Reproduction cueing methodology best discriminated between the groups. Gass advocated that retrieval from memory should be assessed moving forward because “many examinees, including emotionally disturbed and neurologically compromised persons, probably acquire substantially more information than may be implied by measures of free recall” (p. 483).

In their analysis of Verbal Paired Associates, Larrabee and Crook (1995) observed possible problems with the sensitivity of the test. Specifically, they criticized the test as having a low ceiling because it only contained four low-association (i.e.,

"hard") word pairs.

Axelrod, Putnam, Woodard, and Adams (1996) noted that the administration time of the WMS-R (45-60 minutes) was time-consuming and suggested that many clinicians had resorted to shortening the administration time by giving specific subtests (e.g., LM, VR, and VPA) rather than administering the entire test. They developed prorated equations for the General Memory and Delayed Recall indices in 1995, and in 1996 they cross-referenced their equations on a sample of 262 suspected TBI patients. Results of the study found that almost all patients (i.e., 92%) obtained prorated scores that fell within six points of their actual scores. Interestingly, they did not provide data on how much time could potentially be saved by using their methodology. In 1997, (Hoffman, Tremont, G Scott, Adams, & Mittenberg) followed-up on the work on Woodward and Axelrod (1995) by successfully cross-validating their equations on data from an earlier study that included a sample of closed-head injured patients aged 25 to 34 (Mittenberg, B. Burton, Darrow, & B. Thompson, 1992).

In 1999, Golden, White, Combs, Morgan, and McLane noted the criticism of WMS-R had resulted in the creation of new memory tests, including the Memory Assessment Scale (Williams, 1991). Golden and colleagues noted that the research had been inconsistent regarding the relationship between the Memory Assessment Scale and the WMS-R, and they examined the issue using a sample of 51 inpatient neurology participants who had been given both tests. The sample was older (mean age = 55.26, SD = 20.94) but representative for education (M = 11.94, SD = 3.18) and sex (59% male).

Intertest correlations were calculated and reported. However, Golden and colleagues noted, "such correlations are generally underestimates of the actual

correlations between the underlying constructs. This can be addressed statistically […] the resultant correlation […] may better reflect the correlations between the constructs which are being measured" (p.269). Partial correlations were calculated to assess for effects of extraneous influences, such as age, sex, and intelligence and found to be "minimal" (p. 269.) The authors found that the relationship between domains was weaker than would be expected if the tests were measuring the same constructs, and they noted, "These scores are not interchangeable and cannot be used to predict one another" (p. 269). They went on to suggest that the indices, including general memory, do not measure what they were designed to measure and can result in misleading interpretations, even wrong predictions about memory ability and prognoses. Golden and colleagues recommended administering a variety of memory tests to assess a range of abilities to compensate for the different abilities measured by the tests

The WMS-R was designed to address the shortcomings of the WMS, which had been in widespread use for over 40 years by the time the WMS-R was published. Authors of the WMS-R attempted to improve the test by improving the standardization process, publishing data about relatability in the test manual, describing the structure and psychometric properties of the test, and adding new subtests to address gaps in the WMS. The WMS-R also added delayed recall conditions, which had been lacking in the original version of the WMS.

The WMS-R was well-received by clinicians and widely used. Unfortunately, its improvements over the WMS did little to silence its critics, which were numerous. Important shortcomings included a lack of information decay on the Delayed Memory Index (Powel, 1988), a complicated weighting system for calculating index scores, a

small standardization sample (Elwood, 1991), insufficient methods for controlling for intellectual deficits and other cognitive confounds in the standardization sample, and no assessment of performance validity in the standardization sample.

Further, the factor structure was found to be vastly different than advertised in the test manual, including a lack of ability to distinguish between verbal and visual memory problems (e.g., D. B. Burton et al., 1993; Loring et al., 1989; and Roh et al., 1990), no ability to distinguish between errors in storage versus errors in retrieval (Gass, 1995), and poor reliability and validity when compared to other memory tests (e.g., Golden et al., 1999). The WMS-R also made no mention that visual information can be encoded, stored, and retrieval verbally, serious potential confounds for assessing "visual" memory. Finally, the WMS-R, like the WMS, was not linked to any explicit theory of memory, which is unfortunate because our understanding of the neuroanatomy of memory had increased substantially since the WMS was published (e.g., Loring, 1989).

Wechsler Memory Scale - Third Edition (WMS-III)

The Wechsler Memory Scale – Third Edition (WMS-III; Wechsler, 1997) was published only ten years after the WMS-R. The test publishers, possibly hoping to avoid the backslash from the scientific community that followed release of the WMS-R, consulted with experts, focus groups, practicing psychologists, and others when planning and developing the WMS-III. In reading the test manual, it seems they were especially aware of the criticisms of Loring (1989) for example, regarding the lack of attention to scientific advances since the publication of the WMS (e.g., Figural Memory and Visual Paired Associates were dropped).

While their review was limited to the information available from the WAIS-III/WMS-III standardization studies, the review by Horton and Larrabee (1999) is relevant given the sheer amount of information the WMS-III publishers provided for test consumer. Horton and Larrabee (1999) noted that there were numerous other important changes from WMS-R to WMS-III (Wechsler, 1997). A second story was added to Logical Memory to replace WMS-R Story B. The story was administered twice to assess learning of trials. The scoring was also changed to include thematic content. A recognition trail was added after delayed free recall. Visual Reproduction was retained as an optional subtest, though the designs were changed. Trials assessing recognition and copying ability were added. Mental Control was retained as an optional subtest, and more items were added. Verbal Paired Associates was changed drastically from WMS-R to WMS-III. Eight new low associated word pairs were added (i.e., "hard" pairs). More importantly, all high-association (i.e., "easy" pairs) were removed from the test. For Spatial Span, a new stimulus card was developed to make administration and scoring easier.

New visual memory subtests, Faces and Family Pictures, were developed, possibly in response to Loring's (1989) criticisms. Faces is a recognition memory test for faces, and Family Pictures measures recall and recognition of complex visual information. Word Lists, another new subtest, is a list learning test involving 12 words presented over four trials, followed by an interference trial and short- and long-delay trails. A new working memory test, Letter-Number Sequencing (LNS), was added. LNS is like a complex Digit Span test; the examinee is read a series of numbers and letters and asked to rearrange and repeat them in numerical order, then alphabetical order.

WMS-III also featured a new subtest and index structure. Importantly, the General Memory Index on WMS-R measures immediate recall, but the same index on the WMS-III assessed delayed recall. The core test consists of six subtests: three auditory memory subtests (LM I and II, VPA I and II, and LNS) and three visual memory subtests (Faces I and II, Spatial Span, and Family Pictures I and II). The six subtests are combined to produce eight indices: (a) Auditory Immediate (LM I, VPA I), (b) Visual Immediate (Faces I and Family Pictures I), (c) immediate memory (LM I, VPA I, Faces I, and Family Pictures I), (d) Auditory Delayed (LM II and VPA II; (e) Visual Delayed (Faces II and Family Pictures II), (f) Auditory Recognition Delayed (recognition scores for LM II and VPA II; (g) General Memory (LM II, VPA II, Faces II, and Family Pictures II), and (h) Working Memory (Spatial Span and LNS). Information and Orientation were included as optional subtests. WMS-III administration time (i.e., 35 minutes of testing, with a 25- to 30-minute delay in between LM I and LM II) was similar to WMS-R (Wechsler, 1997).

The normative sample for WMS-III included 1250 participants aged 16 to 89. This was a significant improvement over the WMS-R, which included only 316 participants. Because the WMS-III was co-normed with the Wechsler Adult Intelligence Scale – III (WAIS-III; (D. Wechsler, 1997b), comparisons could be made between memory and intelligence test scores. A technical manual was also published, which for the first time provided a theoretical discussion of intelligence and memory in the context of the WAIS-III and WMS-III (D. Wechsler, 1997a).

As described by Horton and Larrabee (1999), the reliability statistics found in the technical manual showed generally superior performance over the WMS-R. For example,

WMS-III median internal consistency reliability for subtests contributing to the Primary Indexes was .81, and reliability for the primary indexes was .87. WMS-R was combined subtest and index reliability was .74. Horton and reliability for the primary indexes was .87 for the WMS-III and Larrabee also reported better test-retest reliability statistics for WMS-III subtests (.62 to .82) and indexes (.70 to .88) than for WMS-R subtests (not reported) and indexes (.57 to .93).

The validity studies in the WAIS-III WMS-III Technical Manual represented a tremendous increase in attention to validity over earlier versions of the test and serious attempts by the test publishers to assess the WMS-III and place in it a firm context with other neuropsychological tests. Results of all the studies are too numerous to list exhaustively (the validity chapter is 104 pages), but the highlights include correlational data between the WMS-III with many other tests, including WMS-R; Wechsler Individual Achievement Test (WIAT); WAIS-III; WAIS-R; Dementia Rating Scale (DRS); Trail Making Test; California Verbal Learning Test (CVLT); Rey-Osterrieth Complex Figure; Boston Naming Test; Judgment of Line Orientation Test; Wisconsin Card Sorting Test; Finger Tapping Test, and Grooved Pegboard Test. As described by Horton and Larrabee (1999), the highest correlations were found between WMS-III and other memory tests and the lowest were between WMS-III and motor tests and the WCST. The highest correlations between WMS-III and WAIS-III were between the two working memory indexes, while the lowest were between the WAIS-III WMI and WMS-III Visual Immediate Recall Index.

The WMS-III publishers also conducted validity research looking at specific patient populations (N = 104), including Alzheimer's disease, TBI, temporal lobectomy,

chronic alcohol use disorder, Huntington's disease, Korsakoff's disease, attention-deficit/hyperactivity disorder (ADHD), learning disorder (LD), multiple sclerosis (MS), Parkinson's disease (PD), and toxin exposure. As noted by Horton and Larrabee (1999), the relationships were smaller in magnitude than with the standardization sample for VIQ and WMS-III Auditory Immediate Memory Index (.38 vs. .58, respectively) and for WMS-III General Memory Index (.32 vs. 56, respectively).

The technical manual's review of the factor structure of the WMS-III is much more extensive than the research conducted with WMS-R. Five models were analyzed, and the best solution across three different age groups was a five-factor model composed of (a) Attention/Concentration, (b) Auditory Immediate, (c) Visual Immediate, (d) Auditory Delayed, and (e) Visual Delayed.

Overall, Horton and Larrabee's review was positive and reflected the tremendous attention the WMS-III development team paid to the standardization methodologies and validity studies during the norming process. The one criticism regarded the lack of a factor analytic study combining WAIS-III and WMS-III because earlier research investigating the factor structure of the WAIS-R and WMS-R (see above) found poor support for the validity of the visual memory subtests and for WMS-R Spatial Span.

In their review of the WMS-III, Tulsky, Chiaravalloti, Palmer, and Chelune (2003) stated that four ideas or conceptual shifts occurred from prior versions of WMS. First, several components of memory were to be assessed, including encoding, storage, and retrieval, and new tests and indices were developed to differentiate among these processes. Immediate and delayed recall indices were retained with the WMS-III, and recognition memory assessment was added to assess for differences in storage and

retrieval (see Gass, 1995; Loring, 1989).

As noted by (Tulsky et al., 2003) the second conceptual change introduced in WMS-III included the introduction of process scores, reflecting the growing popularity of the process approach to neuropsychological test interpretation (e.g., Kaplan, 1988). Tulsky and colleagues noted that despite their popularity, the process scores should be viewed exploratory and taken less seriously than the core index scores because these scores have poor reliability and are not from a standardized distribution, and there is less clinical and research background which to make firm conclusions about what "impaired" scores really mean (p. 105).

A third conceptual shift noted by Tulsky and colleagues (2003) for the WMS-III included a focus on ecological validity, and test authors made a concentrated effort to create tests that would be representative of tasks examinees encounter on a daily basis in real life (this shift is in contrast to earlier versions of WMS, where the goal was to create "pure" versions of tests - for example, the development of abstract designs that could not be verbally encoded). As a result, tasks on WMS-III were designed to mimic real world activities, such as remembering a news report (Logical Memory Story B) or faces of people (Faces subtest). Test developers focused more on how the information was presented rather than on making assumptions about how the brain might process specific types of stimuli. One consequence of a refocus on presentation type is that verbal material was renamed as "auditory," reflecting the format of presentation. This emphasis was also retained in the WMS-IV.

Finally, test developers for WMS-III renamed the Attention/Concentration factor from prior WMS and renamed it Working Memory (Wechsler, 1997). The goal was to

create a parallel with the WAIS-III Working Memory Index. Other changes to WMS-III noted by Tulsky and colleagues included a larger, more representative sample, with stratification based on age range, education level, ethnicity, and sex. While their review is favorable, it was not lost on the author that the editors of the book chapter also served on the WMS-III advisory board.

Overall, however, early reviews of the WMS-III were generally positive, especially in comparison to those of the WMS-R. Many focused-on replicating or expanding on the validity studies published in the technical manual. For example, a study by Mahrou, Devaraju-Backhaus, Espe-Pfeifer, Dornheim, and Golden (2000) examined the relationship between the WMS-III and the WCST. As the reader may recall, these relationships were amongst the weakest noted in the WAIS-III WMS-III Technical Manual. Mahrou et al. (2000) suggested that the WCST required similar abilities as the Working Memory Index (WMI) from the WMS-III. They examine the relationship using a clinical sample of 41 outpatients referred for neuropsychological evaluation. The sample was middle-aged overall ($M = 36.90$ years, $SD = 15.00$) and highly educated ($M = 14.38$ years, $SD = 6.83$). Reported correlations were significant at $p < 0.01$. WCST variables were similarly related to most WMS-III indexes, with similar results for General Memory Index (GMI; 0.42 to -0.56); Auditory Immediate Memory Index (AIMI; 0.40 to -0.56), Visual Immediate Memory Index (VIMI; 0.45 to -0.57. WMI correlations with WCST variables ranged from 0.60 to -0.52, and Auditory Delayed Index (ADI; 0.45 to -0.55.). Interestingly, moderate negative correlations were found for and some WCST variables and Visual Delayed Index performance (VDI; -0.40 to -0.45) and for Auditory Recognition Index (ARI; -0.39 to -0.52).

In another study using more participants from the same clinical sample, Migoya, Zimmerman, and Golden (2000) performed an exploratory factor analysis to evaluate the structure of the WMS-III principal components analyses with varimax rotation. The results revealed 2-, 3-, and 4-factor solutions which accounted for 75%, 83%, and 88% of the common variance, respectively, with the 3-factor solution having the best fit. The first factor (General Memory) had included all subtests and indexes other than VPA II, LNS, and Spatial Span. The second factor (Verbal Memory) included auditory subtests and indexes, and the third factor (Working Memory) included loadings from LNS and Spatial Span. The authors noted that they failed to find support for other factors identified by prior research, visual memory, immediate memory, and delayed memory and that their results could have been due to sample differences, and indeed their sample size of 81 would have been considered "very poor" for PCA by published guidelines (Comfrey & Lee, 1992).

A study by Basso, Harrington, Matson, and Lowery (2000) examined sex differences on VPA and Faces subtests and in a sample of 26 male and 26 female undergraduate students. Results for VPA showed that women had higher age-corrected scaled scores on Total Recall Across Trails ($F(1, 49) = 6.93$, $p = .01$), First Trial Recall ($F(1,49) = 5.03$, $p = .02$), and Percent Retention ($F(1, 49) = 3.80$, $p = .05$) indices. Effect sizes ranged from modest to moderate. IQ did not account for significant proportion of variance on any VPA subtest. It is notable that no statistically significant difference was found between the sexes for VPA Delayed Recall. Overall, men and poor free recall across the four learning trials and had worse retention with a difference of about 2 scaled score points. Recognition memory performance, however, was similar for men and

women. Further, no significant differences were found on any of the Faces subtests. However, higher IQ was related to better recall on Faces I ($F(1,49) = 4.99$, $p < .05$). The authors concluded that the study was limited in its generalizability due to its sample characteristics in the fact that not all WMS-III subtests were administered. Despite these potential limitations, they concluded, "in some instances, the WMS-III norms may result in the erroneous interpretation that men's performances are below expected levels" (p. 234).

Administration time varied considerably from the WMS (15 minutes) to the WMS-R (45-60 minutes). As noted earlier, concerns about time limitations led some clinicians and researchers to use prorated equations on the WMS-R, with generally good results (Axelrod et al., 1996). In 2001, Axelrod conducted a study examining subtest completion times for the WAIS-III and the WMS-III 81 veterans referred for neuropsychological evaluation. The sample was middle-aged ($M = 48.7$, $SD = 14.1$) years, with a mean education of 12.10 ($SD = 2.30$). For WAIS-III subtests, Block Design required the longest administration time ($M = 10.4$ minutes, $SD = 2.9$), though most subtests required less than 5 minutes to administer. For WMS-III subtests, VPA I required the longest administration time ($M = 6.00$ minutes, $SD = 1.4$), though most subtests required 1 to 5 minutes to administer. Administration time and performance were significantly related for some WAIS-III subtests and indexes, but the only significant WMS-III relationship was administration time and performance on the WMI. Overall, the WMS-III was found to take 42 minutes to administer, on average, longer than the time reported in the WMS-III manual (Wechsler, 1997c).

In 2000, Axelrod and Woodard developed three equations for prorating WMS-III

Index scores using a VA sample of 252 clinical cases. The combination of LM + VPA with either Faces or Family Pictures resulted in estimated scores that accounted for 95%+ of the variance in Immediate and General Memory. 80%+ of the estimated scores fell within 3 points of actual sum of scaled scores. The combination of LM + VPA predicted 87% of the variance in scores but only 60% of estimated scores fell within 3 points of actual sum of scaled scores. In a follow-up study similar to the one by Axelrod et al. (1996), Axelrod, D. Dingell, Ryan, and L. Woodard (2001) examined the ability of prorated scores to predict WMS-III performance in a sample of 214 veterans referred for neuropsychological evaluation and a VA hospital. Sample was middle-aged (M = 21.70 years, SD = 13.00), with 12.5 (SD = 2.20) years of education. 44% had been diagnosed with a one or more substance use disorders, 28% with psychiatric disorders, and 2% had no mental health diagnosis. Six equations were tested for their ability to predict the standard WMS-III indices. By eliminating either Faces or Family Pictures, the authors found a time savings of approximately 20% was possible. A savings of more than 50% was possible using two-subtest prorated forms, and calculation of both Immediate Memory and General Memory scores could be accomplished in as little as 20 minutes. Overall, the results were similar to the results of the initial validation study (Axelrod & Woodard, 2000), with 95%+ of scores falling within two standard error of measurement of the full WMS-III indices. The authors cautioned however, that very high or very low scores could produce “less stable” results using their equations, but they noted that 80-90% of the cases in their study had estimated WMS-III scores that fell within 4 points of their actual score.

Not all reviews of the WMS-III were positive, however. For example, in

reviewing the reliability of the WMS-III using the WAIS-III WMS-III Technical Manual, Iverson (2001) noted that for clinical populations (i.e., Alzheimer's disease, chronic alcohol abuse, and schizophrenia) high reliabilities were found for Auditory Immediate Index (AII), IMI, ADI, and the GMI. The most reliable subtests were LM I and VPA I. Iverson noted that the other WMS-III subtests "do not have high reliability" as defined by low internal consistency (< 0.80) and low test-retest reliability (<0.70), or test-retest reliability (<0.60) (p.185). Regarding change in multiple test scores over time, such as with the WMS-III, Iverson recommended using the standard error of the difference over standard errors of measurement (good only for single test scores) or clinical judgment (good only if one is feeling lucky). Iverson described a method for determining reliable change over time using this procedure. Unfortunately, Iverson's procedures were based on small sample sizes in which the disciplines were not retested. As a result, he conceded, there was no way to account for unintended artifacts in his results (e.g., regression to the mean). He concluded by advising for additional research examining specific populations to better predict test-retest reliabilities with the WMS-III.

Millis, Malina, Bowers, and Ricker (1999) offered early criticism for the publisher's proposed factor structure of the WMS-III. In examining the 11 subtests for the standardization sample using confirmatory factor analysis (CFA), they found that a three-factor solution (i.e., working memory, visual memory, and auditory memory) best fit the data. Importantly, while their model did not support that of the publisher's, it did at least offer some support that the WMS-III measures visual memory, something that almost no one believed the WMS-R or WMS accomplished. Unfortunately, the authors were highly critical of the visual memory factor overall, referred to it as "quite flawed,"

and noted “the Faces subtest appears to have insufficient commonality with Family Pictures” (p. 91). They noted this was a serious problem because no other subtest could be substituted for Faces, which was a primary subtest. Further, the authors failed to find support for separate immediate and delayed recall factors.

In another CFA study with the WMS-III, Price, Tulsky, Millis, and Weiss (2002) also failed to find support for the proposed five-factor model using the data from the standardization sample. Like Mills and colleagues (1999), they found a three-factor model best fit the data using CFA and structural equation modeling that included working memory and, for the first time, immediate and delayed contributions to the factors verbal memory and visual memory. The authors noted that their results supported the three-factor structure proposed by Mills and colleagues (1999). They also noted, like Mills and colleagues, that immediate and delayed factors were separate, and they attributed this to the covariance in scores due to the sample characteristics of the standardization sample (i.e., normals had similar performance on immediate and delayed subtests because their memory functioning was intact). They suggested that separate factors might emerge in clinical samples.

In 2003, Tulsky and Price attempted to address the structural discrepancies by developing a six-factor model of cognitive functioning by including subtests from both the WAIS-III and WAIS-IV with the goal of “developing a single battery measuring an integrated model of cognitive functioning across the WAIS-III and WMS-III” (p.149). Using CFA with structural equation modeling, they found that a six-factor model (i.e., verbal comprehension, perceptual organization, auditory memory, visual memory, working memory, processing speed, associates, and sequencing) best fit the WAIS-III

and WMS-III standardization sample data for the 26 subtests included in the study. In a follow-up study, CFA of the models resulted in significant changes, which resulted in improved goodness of fit. Most importantly, tests found to load on multiple factors were removed, including Picture Arrangement (PA), Arithmetic (AR), Spatial Span, and VR I and II.

In a final study, the authors cross-validated the six-factor model is an independent validity sample consisting of 828 examinees who completed the WAIS-III and WMS-III and who met criteria for inclusion in the standardization sample. Age range was 16 to 88 years (M = 36.5; SD = 21.7), and FSIQ was average (M = 95.50, SD = 21.70). Results of the CFA revealed results similar to that of the initial sample and supported the six-factor model. These results, like Mills and colleagues (1999), did not find support for separate immediate and delayed memory factors. The results also replicated those of Mills and colleagues (1999) regarding the differences between the Faces subtest and other visual memory subtests. To address this issue Tulsky, Ivnik, Price, and Wilkins (2003) developed advised replacing Faces with VR and developed norms for the combination of VR and Family Pictures.

As described above, in a CFA study with the WMS-R D. B. Burton et al. (1993) noted that the factors involved included verbal memory, nonverbal memory, attention, immediate, and delayed recall. In study with a similar methodology, D.B. Burton, Ryan, Axelrod, Schellenberger, and Richards (2003) performed a CFA on the WMS-III standardization data to assess construct validity of seven structural models, including those published in the WMS-III manual. As the study was exceptionally thorough and well received, it will be described in some detail. The sample consisted of 281 veterans

evaluated for suspected neuropathology who were divided into three age groups, as well as the 1,250 participants in the standardization sample. It is noteworthy that the sample was 96% male, with a mean age of 51.90 years (SD = 14.50). Mean WAIS-III FSIQ was lower than the standardization sample (M = 88.90, SD = 14.70), and obtained WMS-III index scores were also lower than the standardization sample. IMI, GMI, and WMI was 81.07 (SD = 16.34), 84.82 (SD = 16.26), and 88.66 (SD = 14.69), respectively. 7% have been diagnosed with cardiovascular disease, 9% with TBI, 4% with epilepsy, 7% with dementia, 38% were substance use disorder, 28% with psychiatric disorder, and 4% with no diagnosis. The intercorrelation matrix used for cross-validation of results was taken from the WAIS-III WMS-III Technical Manual.

Overall, the clinical sample means and standard deviations for all 14 WMS-III subtests were approximately one standard deviation lower (e.g., VPA II M = 7.29, SD = 3.62) than the mean score for the standardization sample. Correlations for WMS-III subtests in the four samples were subjected to CFA for the seven structural equations described above, and evaluation of the models was accomplished using the Adjusted Goodness of Fit Index (AGFI). Recall from earlier that the WMS-III publishers stated that a five-factor model provided the best fit to the data across the three age bands in the standardization sample (Model VII in this study, which included factors for factors immediate auditory memory, delay auditory memory, immediate visual memory, delay visual memory, working memory, and learning).

Results of the CFA and chi-square analyses showed that the best fitting model was not the model the model suggested by the WMS-III technical manual, but instead a model that divided the general memory factor into an auditory memory factor and visual

memory factor, along with a working memory factor and learning factor. Overall, the authors suggested that a four-factor model best fit the data and was "significantly more accurate in explaining the intersubtest variability of the WMS-III and generally provided a better fit to the data across all four samples" (p. 638).

The authors described the advantages of their model in terms empirical support and in terms of accepted neuroanatomical models of memory funcitoning. Empirically, the authors noted that their best fitting model did not support the immediate versus delayed recall distinction made by the test publishers, possibly because the publishers failed to include the WMS-III supplemental subtests in their analyses. They also hinted that the publishers may have neglected to mention that they included extra parameters in their model to improve its apparent goodness of fit. Regrading greater neuroanatomical support for their model, the authors referenced earlier research suggesting that frontal lesions are associated with declines in working memory and list learning tasks, while temporal lobe pathology causes deficits in story memory, verbal-paired associate learning, and figure reproduction (e.g., Mennemeier et al., 1994; Stuss et al., 1994). They suggested that the WMS-III list learning tasks "provide a measure of the individual's ability to conceptually organize information in a manner that facilitates their auditory and visual declarative memory" (p. 639). They noted that this view was consistent with that of the test publishers, as was their view that the attention factor was the Attention/Concentration factor identified in the WMS-R. Importantly, the authors results also provided support for the publisher's findings that found a connection between auditory memory and the dominant hemisphere and visual memory with the nondominant hemisphere.

The WMS-III represented major improvements over previous versions of the WMS, including a rigorous standardization process (i.e., 1,250 persons based on 1995 U.S. Census data) that attempted to account for effects of education, sex, race, and geographical region. Also, all participants were screened for medical and mental health problems. Further, the age range of the test was expanded from 16 to 74 years to 16 to 89 years. Vast improvements were made in terms of reliability and validity in the form of over 100 pages of research results published in the WAIS-III WMS-III Technical Manual. Unfortunately, no theoretical model of memory was described in the manual, though there was some discussion of the neuroanatomy of memory and its relation to different memory modalities.

Numerous changes to the content of the test (i.e., addition of Faces, Family Pictures, Word Lists, and Letter–Number Sequencing) resulted in 10 primary subtests and 7 optional subtests, which contributed significantly to longer administration times. The changes also resulted in changes to the factor structure of the test, which was a source of considerable debate, like the prior versions of the WMS. The technical manual described a five-factor structure consisting of working memory, auditory immediate memory, visual immediate memory, auditory delayed memory, and visual delayed memory. As described above, however, this factor structure was not supported by initial exploratory factor analysis studies nor by later more sophisticated CFA investigations using structural equation modeling. Fortunately, Corporation (2002) published an update to the technical manual supporting more recent research, which found a three-factor model consisting of auditory memory, visual memory, and working memory.

On the positive side, research generally supported the idea that sex differences are

minor, with females possibly scoring somewhat higher on some aspects of VPA than males. Overall, however, sex differences were negligible. Unfortunately, the problems with visual memory subtests and indices persisted from WMS-R to the WMS-III. The addition of the Faces subtest was unhelpful, as numerous factor analytic studies demonstrated it measured something other than visual memory (or at least a different aspect of visual memory than the other subtests). Other than this modification, little was done to improve visual memory from WMS-R to WMS-III. Finally, the results of factor analytic studies continued to not provide firm support for the idea that immediate and delayed memory were assessed as independent latent factors on the WMS-III, which was also a point of criticism for the WMS-R and the WMS.

Wechsler Memory Scale – Fourth Edition (WMS-IV)

The WMS-IV (Wechsler, 2009b) was co-normed with the WAIS-IV. Like the WMS-III, the test publishers conducted extensive field studies to inform the development of the WMS-IV, including interviewing users of the WMS-III, and an expert panel was again convened to advise its development (Wechsler, 2009). As with WMS-III, information about reliability, validity, factor structure, and clinical utility are provided in a technical manual (D. Wechsler, Pearson Education, & PsychCorp, 2009).

As with prior versions of the WMS, many changes were made to the basic layout of the WMS-IV. The first notable change is that an Older Adult battery was added to address problems such as fatigue and floor and ceiling effects. The administration time for the Older Adult battery is shorter, and several subtests included in the Adult Battery are not included in the Older Adult battery. Finally, California Verbal Learning Test, Second Edition (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000) scores can be

substituted for Verbal Paired Associates, a move presumably made to shorten administration time for clinicians using both instruments.

The index structure of the WMS-IV has been made simpler from WMS-III. GMI was dropped, and recognition memory is no longer included in the DMI. The Adult battery consists of four core memory subtests, each with an immediate, delayed, and recognition trial and two subtests measuring visual working memory. The WMS-IV Older Adult battery has three of the four memory tests and one of the visual working memory tests. There are five core WMS-IV indexes: Immediate, Delayed, Auditory, Visual, and Visual Working Memory.

The technical manual suggests that the factor structure of the WMS-IV is tighter than that of prior versions of the WMS (2009). While co-normed with the WAIS-IV, the two tests have no overlapping content. Unlike prior versions of the WMS-IV the publishers conducted a joint WAIS-WMS factor analysis and identified a seven-factor solution that best fits the data (i.e., Verbal Comprehension, Perceptual Reasoning, Auditory Working Memory, Visual Working Memory, Processing Speed, Auditory Memory, and Visual Memory). Combined WAIS-IV/WMS-IV factors were also identified, including Quantitative Reasoning, Combined Working Memory, General Memory, Retention, and Retrieval. The publishers produced a computerized scoring program called Advanced Clinical Solutions (ACS; 2009a) that provides supplemental information, such as additional scores, effort measures, demographically adjusted norms, reliable change scores, and a test of premorbid functioning (TOPF). The optional WMS-IV Flexible Approach, which is an abbreviated form of the WMS-IV that provides prorated index scores extrapolated from LM and VPA scores, is also scored using the

ACS software.

Standardization of the WMS-IV continued to improve over WMS-III, with larger samples of different age groups and improved screening of cognitive dysfunction (Wechsler, 2009b). While test administration time may be slightly less than that of WMS-III, requiring 45 to 60 minutes to administer the primary subtests, not including the delay, it unfortunately continues to require a large time commitment from examiners and examinees.

As noted by Drozdick, Holdnack, and Hilsabeck (2011) reliability studies in the standardization sample have several weaknesses, including a lack of data about test-retest reliability with clinical populations, only 23 days between testing and retesting for the standardization group, and various problems with the internal reliability of the DE I and II subtests. Validity in clinical groups is uncertain based on data in the technical manual (this was also a problem with WMS-III) because of small sample sizes.

The research examining the psychometric properties, reliability, and validity of the WMS-IV is more sparse than prior editions of the test. If early reviews are credible, the relative decrease in scientific scrutiny could be because the publishers have improved the instrument significantly from WMS-III. For example, in 2011, Hoelzle, Nelson, and Smith published a study comparing the dimensional structure of the WMS-IV to that of the WMS-III on the standardization samples. They noted that the CFA results included in the WMS-IV technical manual do not include both immediate and delayed memory subtests because "correlations among [them] were greater than the correlations among subtests within the same domain," such as LM and VPA (p. 284). The authors used similar methodology as the test publishers to assess whether the WMS-IV had an

improved factor structure as compared to the WMS-III. The authors used exploratory PCA with parallel analysis (PA) to describe the factor structures. Results supported a one- or two-factor solution across all three age ranges, and a two-dimensional structure was observed across all age ranges: auditory learning/memory (LM and VPA) and visual attention/memory (Designs, DE; VR, Spatial Addition, and SSP).

A similar methodological procedure using the WMS-III standardization data revealed support for retention of three components across some, but not all, of the samples (i.e., verbal memory, visual memory, and working memory). Two-dimensional solutions were also not replicable across samples, though the most frequently observed solution included a general memory dimension (LM, VPA) and a facial memory dimension. The authors concluded that two- and three-dimensional factor structures for WMS-III are difficult to characterized because "significant variability across solutions precludes presentation of average pattern matrix loadings" (p. 288).

The authors suggested that the improved structure of WMS-IV is attributable to the inclusion of the new subtests (i.e., DE, Spatial Addition (SA), and SSP). They also noted that their results suggest that the tests do not appear to be verbally mediated, which, if confirmed, would represent a significant improvement over WMS-III. The authors suggested that the dimensions identified could be useful for localizing "modality-specific memory functioning," which seems to reflect, perhaps for the first time in the reviewing the research of the Wechsler Memory Scales, the suggestion that the test might be linked to verifiable neuropathological and/or empirical theories of memory functioning, such as the Cattell-Horn-Carroll cognitive ability framework, which includes separate auditory and visual memory constructs (McGrew, 2009).

The authors also point out the inconsistency in their findings and that of the WMS-IV factor indices (i.e., Auditory Memory, Visual Memory, and Visual Working Memory) – namely that they are incompatible. They noted that the most common three-factor solution across age ranges was visual attention/memory, LM subtests, and VPA subtests. They noted that distinguishing between Visual Memory and Visual Working Memory indices is difficult, though they noted that they did not contain other factors that might suggest they are verbally mediated). Hoelzle et al. (2011) advised against conducting CFA using both immediate and delayed subtests due to the high correlations between the two variables. Instead, they advised that CFA of clinical samples would be interesting to inform whether their three-factor model is superior to the two-factor model described in the test manual. They also suggested that replication of their findings to clinical samples would be useful and that "efforts to determine whether psychometric proprieties of neuropsychological measures are similar across diverse samples with localized or lateralized cerebral dysfunction would only improve clinical assessment."

The most recent comprehensive review of Wechsler's Memory Scales was by Kent (2016), who succinctly outlined the different versions of the WMS through the current edition. While this author does not agree with many of Kent's recommendations for the next version of the WMS (see below), the author did emulate his writing style and included an interim summary between each version of the WMS to facilitate comprehension. Kent's (2016) basic assumption is that the WMS-IV (and prior versions, for that matter) is a poor test because it is not grounded in an explicit theory of memory. He was also critical of the technical manual's lack of data about various clinical groups. He went on to describe an apparent decline in the quality of graduate school education in

clinical psychology (e.g., less history of psychology courses, little attention to reliability and validity), and he stated, “This trend in training […] is alarming and does not bode well for the future of psychological or neuropsychological assessment” (p. 2).

Kent, like other researchers reviewed above, noted that the WMS-IV continues to not support separate factors for immediate and delayed index scores, though he does note that other memory tests, such as the CVLT-II “suffer from the same problem” in factor analytic studies (p. 15). He briefly reviews the research described previously by Hoelzle et al. (2011) described above, followed by a brief review of the changes index and subtest structure of the WMS-IV. He concludes his review of the WMS-IV by stating, “the WMS-4 is the most radical of all the revisions” and suggests that it should not be compared to previous versions of the test. He also noted that the test no longer measures verbal working memory, a point supported by Hoelzle et al. (2011), but not necessarily by the test publishers, who specific describe an Auditory Working Memory factor in their factor analysis findings (Wechsler et al., 2009).

Kent also criticized the WMS-IV for dropping Digit Span and suggests that doing so could result in clinical decision-making errors in patients who present for disability evaluations with complaints of memory problems. He concluded by recommending that the Delayed Memory Index be renamed the intermediate recall (or memory) index; that the next version of the WMS include Digit Span, along with Logical Memory, Verbal Paired Associates, Visual Reproduction, Mental Control, and Personal and Current Information; that the battery be shortened; that the next version of the test be linked to an explicit neuropsychological theory of memory; that the next version of the test include clinical subgroups of at least 50 cases each; that the next version assess prospective

memory; that it include an alternate form; and that it include "measures of effort and test validity."

Purpose of the Study

The purpose of this study was to examine the relationship between auditory episodic memory across two versions of the Wechsler's Memory Scale (i.e., WMS-III and WMS-IV) and various neuropsychological domains including intellectual functioning as assessed by the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; Wechsler 2008), attention as assessed by the Omission and Commission errors from the Conners' Continuous Performance Test-II (CPT-2; Conners, 2000); executive functions using (a) the Stroop Color and Word Test (Stroop; Golden & Freshwater, 2002), (b) Part B from the Trail Making Test (TMT B; Reitan & Wolfson, 1985), and (c) the Category Test (CT; Reitan & Wolfson, 1985).

Hypothesis One

It was hypothesized that the degree of agreement between WMS-III and WMS-IV as determined by scaled scores would be within one point at a rate of 90% or better for VPA3 and VPA4 and that the two tests would correlate at a level of .70 or above.

Justification. Research in the technical manuals of both WMS-III and WMS-IV demonstrate a strong relationship between each other and between other tests of verbal memory (Wechsler, 1997a; Wechsler et al., 2009). While changes to VPA from WMS-III to WMS-IV were significant (e.g., the reintroduction of semantically-related word pairs, four fewer items on the Older Adult battery than the Adult battery), the two tests remain similar in structure and format and so were expected to correlate highly between one

another in a clinical sample as with the standardization sample. Additionally, changes to the standardization process (i.e., better screening methods for excluding persons from the standardization sample with cognitive impairment) for WMS-IV would be expected to mitigate potential within-participant effects due to the structural changes from VPA3 to VPA4. This was demonstrated in nonclinical individuals via reliability studies published in the WMS-IV Technical Manual, and similar results were expected in this diverse clinical sample.

Hypothesis Two

Hypothesis Two predicted that WMS-IV Verbal Paired Associates would show a significantly stronger relationship with attention, intellectual, and executive functioning ability as measured by performance on CPT-2 Commissions and Omissions, WAIS-IV, Trail Making Test Part B, and by the Category Test in a clinical sample than would WMS-III Verbal Paired Associates.

Justification. Clinical participants have been shown to perform less well on WMS-III compared to the WMS-III standardization sample (e.g., Burton et al., 2003), and while this finding was also expected with WMS-IV, the improvements in sampling methodologies and revised structure of WMS-IV VPA were robust and therefore expected to better relate to intellectual and executive functioning abilities in the present clinical sample. Prior research found small to moderate relationships between VPA and measures of executive function (Horton & Larrabee, 1999). However, researchers have yet to investigate the relationship between executive funcitoning and in a clinical sample who completed both WMS-III and WMS-IV VPA, so this research will allow for direct comparisons. Similarly, prior research has investigated the relationship between VPA and

intelligence in the WMS standardization samples (e.g., (Wechsler, 1997a; Wechsler et al., 2009), but research comparing both versions of VPA to the current gold standard in intellectual assessment is lacking. This study addresses that gap, and stronger relationships were expected between WMS-IV VPA and intellectual functioning than with WMS-III VPA because the ceiling effect restricts the range of correlation coefficients.

CHAPTER III
METHOD

Participants

This study utilized archival data managed by Nova Southeastern University's Psychology Services Center – Neuropsychology Assessment Center (NAC). Participants in this study were all between the ages of 18 and 90. For inclusion in the study, participants must have been 18 years of age or older and have completed the WMS-III, WMS-IV, WAIS-IV, CPT-2, Category Test, Stroop Color and Word Test, and the Trail Making Test. Participants included 36 adults, ages 19 to 67 (M = 36; SD = 14.71) with an education range of 8 to 18 years (M = 13.51, SD = 2.27). 58.3 percent were female, and 72 percent were right-handed. Fifty-eight percent were Caucasian, 14 percent were Hispanic, and 11 percent were African American. Primary diagnoses represented in this sample included neurological disorders (47 percent) and psychiatric disorders (19 percent). Fourteen percent of participants were diagnosed with both a neurological and a psychiatric disorder, and 11 percent received no diagnosis or were missing a diagnosis.

Procedures

Data Collection. All data were collected from psychological evaluations of adults referred to the NAC at Nova Southeastern University (NSU). Supervised by licensed clinical neuropsychologists, doctoral-level graduate students administered all assessments as part of comprehensive neuropsychological evaluations. All students completed NSU CITI certification training. Participants were administered approximately 15-25 hours of testing over approximately two months; for the present research, however, only tests

purported to measure the variables of interest were selected. All protocols were checked for administration and scoring accuracy by advanced graduate students or a licensed clinical neuropsychologist.

Institutional Review Board Requirement

Approval was obtained from the Institutional Review Board (IRB) at NSU to conduct archival research following the approval of the proposed project by the dissertation committee. As mandated by the IRB, all data were de-identified to maintain confidentiality.

Measures

Standardized scores are were used for each of the tests, including T-scores (mean of 50, standard deviation of 10) and standard scores (mean of 100, standard deviation of 15), and scaled scores (mean of 10, standard deviation of 3). Measures were included were those measuring memory, intellectual abilities, working memory, sustained attention, and executive functions as described below.

Category Test. The Category Test (DeFillippis, 1992) consists of seven subtests that involve a series of images that suggest a number from one to four. The first subtest requires the examinee to recognize Roman numerals ranging from one to four (I, II, III, IV). The second subtest requires the examinee to count the number of objects on the computer screen. For subtests 3 through 6, the number is suggested by the spatial location, the orientation of an odd or specific item, or through proportional reasoning. The final subtest is a memory test made up of items administered to the examinee in subtests one through six. The Category Test requires the examinee to determine the

correct strategy to use in each subtest by trial-and-error, as the "rule" remains the same within each subtest. For each item, the examinee is allowed one response; a bell sound indicates a correct response, and a buzzer sound indicates an incorrect response. This feedback prompts the examinee to alter responses until the appropriate "rule" is discovered, which can then be applied to obtain correct responses to the rest of the items in that subtest. The examinee's score is determined by the number of errors the individual makes on the seven subtests (Golden, Espe-Pfeifer, & Wachsler-Felder, 2000). The clinical utility of the Category Test is strong, and performance on the Category Test is one of the best predictors of overall brain dysfunction of all neuropsychological tests (Anthony, Heaton, & Lehman, 1980; Reitan & Wolfson, 1992) because it is sensitive to overall brain dysfunction, rather than localization or lateralization effects, and duration of brain dysfunction does not affect performance (Sweet & King, 2003).

Conners' Continuous Performance Test II (CPT-2). The CPT-2 (Conners & Staff, 2000) is a computerized test of sustained attention and response inhibition. It requires the examinee to maintain a continuous response set and then inhibit responding when a target is presented. While primary indicated for screening and monitoring the effectiveness of treatment, the CPT-2 is also commonly used with other assessment procedures (e.g., clinical interview) to make diagnostic decisions regarding attentional impairment. Omission errors occur when the examinee fails to respond to a nontarget stimulus (i.e., fails to click the mouse when presented with a letter other than "X"). Excessive omission errors are associated with inattentive behavior. Commission errors occur when the examinee erroneously responds to a target symbol (e.g., clicks the mouse when presented with an "X"). Excessive commission errors are associated with

hyperactive behavior. For this study, CPT-2 Omission and Commission errors were used, as intact attention is a prerequisite for memory functioning. Further, CPT-2 omission and commission errors have been found to be sensitive to the types of inattention seen in persons diagnosed with Attention-Deficit/Hyperactivity Disorder (ADHD) (Epstein et al., 2006; Fasmer et al., 2016).

Stroop Color and Word Test (Stroop). The Stroop (Golden, 1978) measures an individual's ability to attend to a goal and suppress an automatic response for a different response. It measures cognitive flexibility and selective attention. It is commonly used to assess brain damage, particularly in the frontal lobes. Stroop Word performance measures an individual's reading speed and reaches adult levels around age 10. Stroop Color-Word measures an individual's ability to inhibit reading the word; instead, the participant states the color of the ink in which the word is printed. For purposes of this research, the color-word score was used as a measure of executive functioning, as it has been shown to assess mental flexibility and response inhibition (Wecker, Kramer, Wisniewski, Delis, & Kaplan, 2000). Further, poor performance on the Stroop as it has been shown in children, adolescents, and adults with frontal lobe deficits (Golden et al., 2000; Homack & Riccio, 2004).

Trail Making Test, Part B. The Trail Making Test (TMT) is made up of two parts, Trails A and Trails B. Trail Making Test (TMT): The TMT is a measure of attention, speed, and mental flexibility. It requires the examinee to connect, by making pen/pencil lines, 25 encircled numbers randomly arranged on a page in the proper order (Part A) and 25 encircled numbers and letters in alternating order (Part B). Part B was included in the study, as it has consistently been shown to predict cerebral dysfunction

(Bowie & Harvey, 2006; Doehring & Reitan, 1962; Wolfson, 1995). It is a robust measure of executive functioning, specifically, mental flexibility (Crowe, 1998; Kortte, Horner, & Windham, 2002). Further, performance on Part B of the TMT has been associated with activation of frontal brain areas involved with executive functioning, including the left dorsolateral prefrontal cortex, cingulate gyrus, and medial frontal gyrus (Zakzanis, Mraz, & Graham, 2005). Performance deficits have been found in patients with frontal brain lesions (Stuss et al., 2001) and persons diagnosed with mental disorders known to affect executive functioning, including Alzheimer's disease (Amieva et al., 1998), Bipolar I and II disorder (Torrent et al., 2006; Zimmerman, DelBello, Getz, Shear, & Strakowski, 2006), and schizophrenia (Heinrichs & Zakzanis, 1998).

Verbal Paired Associates (VPA). Wechsler first used verbal paired associates to evaluate episodic memory in 1917. He adapted an analogy test developed by Woodworth and Wells (1911) for his master's thesis at Columbia University (Wechsler, 1917). In the original version, Wechsler provided examinees with "preformed associates" (pairs of related words) and "new formed associates" (pairs of unrelated words) to participants with Korsakoff's syndrome. Wechsler's format was maintained through his first published memory scale, the Wechsler Memory Scale and for the revised edition of the scale (WMS-R).

The WMS-III VPA subtest included the presentation of eight unique word pairs across four trials. Recall was measured after each trial and again after a 25-35-minute delay by providing the examinee with the first word of each pair and asking the examinee to provide the second. Recognition was measured after the delayed recall by presenting the examinee with pairs of words and having the examinee state whether he or she saw

the word pair during the learning phase. While the format of the WMS-III VPA subtest was identical to earlier versions of the Wechsler Memory Scales, the test items (word pairs) were changed. The test consists of eight different word pairs, and they are all unrelated. This change was made to present a greater learning challenge to the examinee. The "pre-formed" associates were too easy for the examinees. The number of trials also varied with each test; WMS-R administration required at least three trials up to a maximum of six, while for the WMS-III VPA, four trials were administered to all examinees. A recognition trial was also added, where the examinee was asked to "recognize" target words. However, the examinees found this to be easy, and most healthy persons obtained a perfect score. The WMS-III also added an optional list learning task.

Significant changes were made to VPA with the release of the WMS-IV. The number of word pairs for the Adult battery was increased from 8 to 10 (the Older Adult battery has 10). More "easy" items were added to reduce floor effects. Recognition items were modified to include more difficult foils to reduce ceiling effects. An optional delayed word recall trial was also added. VPA requires the examinee to pay attention to the examiner, to listen to and process unrelated word pairs (receptive language, executive functioning), and to recall and express what was learned (expressive language) both immediately and following a 20-30-minute delay. The words on VPA are at first- to third-grade level. The expressive and receptive language demands are lower on VPA than on the other auditory memory test on the WMS, Logical Memory (LM). Both tests require working memory, auditory attention, hearing acuity, and articulation.

Wechsler Adult Intelligence Test, Fourth Edition (WAIS-IV). The WAIS-IV

(Wechsler, 2008) is the current edition of the Wechsler's popular intelligence test. The WAIS-IV measures global intellectual/cognitive functioning in adolescents and adults ages 16 to 90, through the administration of 10 core subtests, including Block Design, Similarities, Digit Span, Matrix Reasoning, Vocabulary, Arithmetic, Visual Puzzles, Information, Coding, and Symbol Search. In 2003, in recognition that cognitive functioning includes more than what was captured by Performance IQ (PIQ) and Verbal IQ (VIQ), the WAIS-IV dropped PIQ and VIQ for the four-factor model used by the WISC-III. The WAIS-IV, published in 2008, utilizes this model, which shifted the focus of interpretation from the level of subtests to the level of indices. Letter Number Sequencing and Cancellation were added to expand the assessment of working memory and processing speed, respectively. Digit Span was also changed significantly to include digit sequencing items to improve the assessment of working memory. Figure Weights and Visual Puzzles were added to extend the assessment of fluid reasoning.

Wechsler Memory Scale, Third Edition (WMS-III). The Wechsler Memory Scale-Third Edition (WMS-III) was released in 1997. Updates from the WMS-R included an extended age range (from 74 to 89 years), interpolated norms were replaced with sampling for each age group, recognition memory tasks were added, and other steps to improve validity and reduce bias. 1,250 cases were used for the standardization sample, which included 13 age groups ranging from 16 to 89. The WMS-III was conormed with the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III) and the Wechsler Test of Adult Reading (WTAR), which allowed for comparisons to be made across the tests. The WMS-III consisted of 11 primary subtests (Logical Memory, Verbal Paired Associates, Letter-Number Sequencing, Spatial Span, Faces, and Family Pictures) and five optional

subtests. Eight primary index scores were derived from the six primary subtests (Auditory Immediate, Visual Immediate, Immediate Memory, Auditory Delayed, Auditory Recognition Delayed, Visual Delayed, Working Memory, and General Memory). Also, four auditory process composites were derived (Single-Trial Learning, Learning Slope, Retention, and Retrieval). The General Memory Index was comprised of the auditory and visual delayed recall tasks and the auditory recognition tasks.

Wechsler Memory Scale, Fourth Edition (WMS-IV). The WMS-IV (Wechsler, 2009) is the current edition of Wechsler's test of memory functioning in adults. The Adult Battery consists of 7 subtests and is administered to individuals ages 16-69. The Older Adult battery consists of 5 subtests and is administered to individuals ages 65-90. Individuals aged 65-69 may be administered either battery. Subtests include Logical Memory, Verbal Paired Associates, Visual Reproduction, Designs, Spatial Addition, and Symbol Span. Noteworthy changes were made from the WMS-III to the WMS-IV. The optional word list from the WMS-III was dropped and the CVLT-II could be substituted for VPA Immediate, Delayed, and Recognition Indexes. The Faces subtest and the Family Picture subtests were dropped, as was Letter-Number Sequencing, Spatial Span, and Mental Control.

CHAPTER IV
RESULTS

Preliminary Analysis

The Statistical Package for the Social Sciences (SPSS) was utilized for all data analyses in the present study. A scan for missing values was conducted prior to analysis using the Frequencies descriptive statistic procedure with SPSS. Unlikely values (outliers) were assessed by visually inspecting the histogram for each variable. The Kolmogorov-Smirnov test were used to test the assumption that the sample data were drawn from a normally distributed population. Skewness and kurtosis were assessed by inspection of the standard of error for each, which is provided by SPSS. Dividing each value by its standard error provided a result that was compared to a standard of ±1.96, so values within the range were considered acceptable. Descriptive statistics for performance of all participants is provided in Tables 1 through 3. No outliers or missing values were found.

Examination of skewness and kurtosis values revealed that all variables were approximately normally distributed except for Conner's CPT-2 Omissions, Stroop Color-Word, and WMS-IV Designs I and II, which were positively skewed; these distributions exceeded the acceptable values of skewness or kurtosis, which indicate these variables were not normally distributed. The Kolmogorov-Smirnov values for each variable were analyzed to further assess whether each were normally distributed, with a cut-off of greater than .05 used to establish normality.

Variables with values less than less than .05 WMS-III Spatial Span (.005),

Conner's CPT-Omissions (.000), WAIS-IV Visual Puzzles (.000), and WMS-IV Spatial Addition (.000).

Table 1

Descriptive Statistics of the Performance for Wechsler Memory Scale, Third Edition (WMS-III)

	M	*SD*	Skewness	Kurtosis
WMS-III Auditory Immediate Index	97.44	16.96	-.15	-.55
WMS-III Visual Immediate Index	95.69	15.49	.09	-.82
WMS-III Immediate Memory Index	96.03	17.28	-.27	.07
WMS-III Working Memory Index	93.42	15.52	.21	-.55
WMS-III Auditory Delayed Index	100.25	15.57	-.68	.28
WMS-III Visual Delayed Index	99.78	16.16	.36	-.25
WMS-III General Memory Index	100.53	17.11	-.07	.01
WMS-III Logical Memory I	9.81	3.45	-.32	.013
WMS-III Logical Memory II	10.42	3.18	-.49	.20
WMS-III Verbal Paired Associates I	9.36	3.12	-.015	-.62
WMS-III Verbal Paired Associates II	9.75	2.77	-.52	-.88
WMS-III Faces I	9.44	3.02	.58	-.60
WMS-III Faces II	10.47	3.23	.58	-.33
WMS-III Family Pictures I	9.25	3.31	.11	-.64
WMS-III Family Pictures II	9.44	2.82	.20	.22
WMS-III Spatial Span*	8.58	3.37	-.20	-.97

Note. M = mean; SD = Standard Deviation; N=36; * = Removed from subsequent analyses.

Samples with significant departure from normality can affect the robustness of parametric tests that assume normal distributions, which can influence inferences about the population. Therefore, these variables were removed from subsequent analyses. Minor violations to the assumption of normality typically has little impact on the analyses, and all other subtests did not exhibit significant deviation from a normal distribution.

Hypothesis One

Hypothesis One stated that the degree of agreement between Wechsler Memory Scale-III and Wechsler Memory Scale-IV as determined by scaled scores would be within one point at 90% or better for VPA3 and VPA4 and that the two tests would correlate at a level of .70 or above. Results show that degree of agreement within one scaled score point was 41.7% for VPA I across WMS-III and WMS-IV. Degree of agreement for two-scaled scores was 61.1%, and

Table 2

Descriptive Statistics of the Performance for Wechsler Memory Scale, Fourth Edition (WMS-IV)

	M	*SD*	Skewness	Kurtosis
WMS-IV Logical Memory I	10.31	2.94	-.79	.30
WMS-IV Logical Memory II	10.44	3.43	-.84	.62
WMS-IV Verbal Paired Associates I	9.44	3.38	.53	-.11
WMS-IV Verbal Paired Associates II	9.92	3.20	-.53	-.55
WMS-IV Designs I*	8.97	3.25	.46	-.36
WMS-IV Designs II*	9.00	2.70	.49	3.27
WMS-IV Visual Reproduction I	7.94	2.99	-.57	.051
WMS-IV Visual Reproduction II	9.31	2.79	-.51	.58
WMS-IV Spatial Addition*	8.08	2.72	.39	-.83
WMS-IV Symbol Span	8.69	2.48	-.26	-.38
WMS-IV Auditory Memory Index	100.06	17.29	-.75	.33
WMS-IV Visual Memory Index	92.56	15.40	-.31	1.19
WMS-IV Visual Working Memory Index	90.61	13.19	-.01	-.28
WMS-IV Immediate Memory Index	94.19	15.68	-.64	.85
WMS-IV Delayed Memory Index	97.50	16.66	-.86	.85

Note. M = mean; SD = Standard Deviation; N=36; * = Removed from subsequent analyses.

88.9% of VPA I scores fell within three scaled score points. Degree of agreement within one scaled score point was 55.6% for VPA II across WMS-II and WMS-IV.

Degree of agreement for two-scaled scores was 72.2%%, and 88.9% of VPA II scores fell within three scaled score points. Further, the degree of agreement was identical for VPA I and II within three scaled score points. However, more cases fell within one scaled score point on VPA II (55.6%) than for VPA I (41.7%).

The magnitude of WMS-III VPA and WMS-IV VPA relationships was greater than or equal to .70 for all relationships except WMS-III VPA I and WMS-IV VPA II, which was .61. Thus, while the percentage of participants with scaled scores within one point was less than predicted, the magnitude of relationships for VPA subtests with other memory subtests was supported in the predicted direction. Overall, the degree of agreement for VPA I and II was lower than predicted. The magnitude of WMS-III VPA and WMS-IV VPA relationships was .76, greater than .70 as predicted. As a result, the hypothesis was partially supported.

Hypothesis Two

Hypothesis Two stated that WMS-IV VPA would show a significantly stronger relationship and thus be better able to predict attention, intellectual, and executive functioning ability as measured by performance on the CPT-2 Commissions, WAIS-IV, TMT B, Stroop Interference, and by the Category Test in a clinical sample than would Wechsler Memory Scale-III Verbal Paired Associates. Table 4 shows correlations for both versions of VPA for all measures.

Correlations were considered significant at the 0.01 level. WMS-III VPA I was significantly related to WAIS-IV Similarities (r = 0.45), WAIS-IV Vocabulary (r = 0.55), and WAIS-IV Verbal Comprehension Index (r = 0.53). WMS-IV VPA I was significantly

related to WAIS-IV Similarities (r = 0.48), WAIS-IV Verbal Comprehension Index (r = 0.44).

Table 3

Descriptive Statistics for Intellectual, Executive Functioning, and Attention Tests

	M	*SD*	Skewness	Kurtosis
WAIS-IV Block Design	7.91	3.45	.90	.72
WAIS-IV Similarities	9.00	4.08	.93	.67
WAIS-IV Digit Span	9.06	3.27	.47	1.33
WAIS-IV Matrix Reasoning	9.46	3.10	-.62	-.31
WAIS-IV Vocabulary	10.31	3.43	.59	.75
WAIS-IV Arithmetic	7.69	2.73	.78	.12
WAIS-IV Symbol Search	8.66	3.22	-.03	-.44
WAIS-IV Visual Puzzles*	9.11	3.22	.62	-.68
WAIS-IV Information	10.03	2.63	-.51	.14
WAIS-IV Coding	8.86	2.83	.33	-.07
WAIS-IV VCI	98.53	15.06	-.33	.36
WAIS-IV PRI	93.38	16.57	.15	-.46
WAIS-IV WMI	91.47	14.76	.03	-.50
WAIS-IV PSI	93.00	15.14	.20	.06
WAIS-IV Full Scale IQ	93.44	14.79	-.67	.37
Trail Making Test Part B	38.83	12.28	-.01	.24
Stroop Color-Word*	45.74	7.78	.482	2.206
Stroop Interference	47.26	8.23	.634	.013
Conner's CPT-II Omissions*	53.99	15.03	2.03	.78
Conner's CPT-II Commissions	54.95	12.42	.36	-.87

Note. M = mean; *SD* = Standard Deviation; N=36; * = Removed from subsequent analyses.

WMS-III VPA II was significantly related to WAIS-IV Similarities (r = 0.49), WAIS-IV Matrix Reasoning (r = 0.48), WAIS-IV Vocabulary (r = 0.50), WAIS-IV Information (r = 0.59), WAIS-IV Verbal Comprehension Index (r = 0.61), and WAIS-IV

Full Scale IQ (r = 0.49).

Table 4

Pearson's Correlation for WMS-III and WMS-IV Verbal Paired Associates and Measures of Intelligence, Attention, and Executive Functioning

	WMS-III VPA I	WMS-IV VPA I	WMS-III VPA II	WMS-IV VPA II
WAIS-IV Block Design	.02	-.15	.13	-.06
WAIS-IV Similarities	.45*	.48*	.49*	.36
WAIS-IV Digit Span	.31	.29	.31	.23
WAIS-IV Matrix Reasoning	.25	.40	.48*	.38
WAIS-IV Vocabulary	.55*	.35	.51*	.31
WAIS-IV Arithmetic	.05	.16	.06	.09
WAIS-IV Symbol Search	.21	.21	.33	.12
WAIS-IV Information	.41	.41	.59*	.50*
WAIS-IV Coding	.27	.40	.28	.24
WAIS-IV VCI	.53*	.44*	.61*	.43*
WAIS-IV PRI	.21	.18	.31	.13
WAIS-IV WMI	.23	.26	.24	.20
WAIS-IV PSI	.25	.31	.32	.18
WAIS-IV Full Scale IQ	.41	.37	.49*	.31
CPT-2 Commissions	.07	.24	.15	.15
Category Test	.26	.23	.28	.15
TMT B	.19	.22	.24	.09
Stroop Interference	-.16	-.09	.01	-.06

Note. * = Correlation is significant at the 0.01 level.

WMS-IV VPA II was significantly related to WAIS-IV Information (r = 0.50) and WAIS-IV Verbal Comprehension Index (r = 0.43).

Using the methodology described by (Lee & Preacher, 2013), each correlation was converted to a z-score using Fisher's r-to-z transformation. Steiger (1980) techniques were then used to compute the asymptotic covariance of the estimates, and these quantities are used in an asymptotic z-test. The standardized scores (i.e., z-scores) show the results of tests of the equality of two correlation coefficients obtained from the

Table 5

Comparisons of Equality for WMS-III VPA I and WMS-IV VPA I for Intellectual, Attentional, and Executive Functioning Ability

	WMS-III VPA I	WMS-III VPA I	Z-score
WAIS-IV Block Design	.02	-.15	1.27
WAIS-IV Similarities	.45	.48	-0.26
WAIS-IV Digit Span	.31	.29	0.16
WAIS-IV Matrix Reasoning	.25	.40	-1.20
WAIS-IV Vocabulary	.55	.35	1.72
WAIS-IV Arithmetic	.05	.16	-0.82
WAIS-IV Symbol Search	.21	.21	0
WAIS-IV Information	.41	.41	0
WAIS-IV Coding	.27	.40	-1.04
WAIS-IV VCI	.53	.44	0.78
WAIS-IV PRI	.21	.18	0.23
WAIS-IV WMI	.23	.26	-0.23
WAIS-IV PSI	.25	.31	-0.47
WAIS-IV Full Scale IQ	.41	.37	0.33
CPT-2 Commissions	.07	.24	-1.29
Category Test	.26	.23	0.23
TMT B	.19	.22	-0.23
Stroop Interference	-.16	-.09	-0.52

clinical sample with the two correlations sharing one variable in common. Using $p < .01$ to determine significance, the results are presented in Tables 5 and 6.

Table 5 shows that no significant differences were found between WMS-III VPA I and WMS-IV VPA I in terms of their ability to predict intellectual, attentional, or executive functioning abilities.

Similarly, Table 6 shows no significant differences were found between WMS-III VPA II and WMS-IV VPA II in terms of their ability to predict intellectual, attentional,

or executive functioning abilities. As a result, Hypothesis Two was not supported.

Table 6

Comparisons of Equality for WMS-III VPA II and WMS-IV VPA II for Intellectual, Attentional, and Executive Functioning Ability

	WMS-III VPA II	WMS-IV VPA II	Z-score
WAIS-IV Block Design	.13	-.06	1.59
WAIS-IV Similarities	.49	.36	1.22
WAIS-IV Digit Span	.31	.23	0.69
WAIS-IV Matrix Reasoning	.48	.38	0.94
WAIS-IV Vocabulary	.51	.31	1.87
WAIS-IV Arithmetic	.06	.09	-0.25
WAIS-IV Symbol Search	.33	.12	1.81
WAIS-IV Information	.59	.50	0.92
WAIS-IV Coding	.28	.24	0.35
WAIS-IV VCI	.61	.43	1.82
WAIS-IV PRI	.31	.13	1.55
WAIS-IV WMI	.24	.20	0.34
WAIS-IV PSI	.32	.18	1.21
WAIS-IV Full Scale IQ	.49	.31	1.70
CPT-2 Commissions	.15	.15	0
Category Test	.28	.15	1.11
TMT B	.24	.09	1.27
Stroop Interference	.01	-.06	0.58

CHAPTER V
DISCUSSION

Historically, there has been a disconnect between the neuroscience of memory (e.g., theory, neuroanatomy) and the formal assessment of memory by neuropsychologists (e.g., assessment and prediction of functioning). The current study examined whether changes to the Verbal Paired Associates (VPA) subtest from the WMS-III to WMS-IV resulted in changes to the way in which the two tests assess memory. The purpose was to examine changes to VPA from WMS-III to WMS-IV in a clinical sample to understand the differences between versions and to identify practical implications for neuropsychologists who use VPA to make decisions about current and future memory functioning.

Hypothesis One

Hypothesis One predicted that the degree of agreement between WMS-III and WMS-IV as determined by scaled scores would be within one point at a rate of 90% or better for VPA3 and VPA4 and that the two tests would correlate at a level of .70 or above. The hypothesis was partially supported by the current analysis.

This hypothesis was proposed because it is unknown how much (if at all) the changes from VPA3 to VPA4 effect the measurement of memory in clinical populations. If the WMS were based on a unified theory of memory, then this process would be straightforward because such a theory would allow for testable hypotheses. However, since the WMS has always been atheoretical, there is no empirical foundation to rest justification of changes from one version of the test to the next, and it was essentially left

to researchers and practitioners to determine this after the test was published for most clinical groups. While clinical subsamples were included the WMS-III technical manual, but the sample sizes were small and otherwise limited. Rationale for the changes provided in the WMS-IV Technical Manual included "inadequate floor at some ages and [...] insufficient data points on delayed recall to have a strong scaled score metric" (Wechsler, 2009, p. 9).

The first part of Hypothesis One was not supported as predicted. While 90%+ of scaled scores were expected to be within one point from VPA3 to VPA4, this was not found to be the case. Current results found that only 41% of participants had a one-point degree of agreement for VPA I, and only 55% had a one-point degree of agreement for VPA II. This unexpectedly low level of agreement, particularly for VPA I suggests that even though the mean scores are similar, there are performance differences from VPA3 to VPA4.

Because the WMS-IV, unlike WMS-III, was normed on individuals who had been screened for acquired or developmental memory impairment, the clinical sample in this research may be closer in similarity to the WMS-III standardization sample that WMS-IV. If so, overall VPA performance across test versions would be expected to correlate favorably because the increased presence of memory problems in the WMS-III standardization sample and the addition of semantically similar items to VPA4 essentially cancelled each other out for clinical outpatients' scaled scores (i.e., mean scaled scores would be similar from VPA3 to VPA4) but not for changes in scaled score points.

Further, this would be more likely to manifest itself as a function of age differences (and possibly education) in the clinical sample, where younger, higher

educated clinical examinees from the clinical sample perform more like individuals from the WMS-IV normative sample (i.e., less problems with episodic memory and less variance overall) and older, less educated clinical examinees perform more like individuals from the WMS-III normative sample (i.e., more problems with episodic memory and more variance in performance overall). The net outcome would therefore produce the observed results from the current research: while scaled score changes are relatively widespread from VPA3 to VPA4, overall mean scores are nearly identical. If accurate, this admittedly could be a function of the limitations of using mean as a measure of central tendency or the effects of assessing this issue in a relatively heterogenous sample as much as it could result from changes from VPA3 to VPA4.

The second component of Hypothesis One, that WMS-III VPA and WMS-IV VPA would correlate at .70 or higher, was supported by the current research ($r = .76$). The results were consistent with the WMS-IV validation studies on the normative and clinical population subgroups. For instance, the correlation for the normative sample of VPA4 I and II was .84 and .85 for the Adult and Older Adult groups, respectively. Scores for VPA I were similar for WMS-III and WMS-IV ($m = 9.36$, $SD = 3.12$ and $m = 9.44$, $SD = 3.38$, respectively). This suggests as a group, outpatient clinical neuropsychology patients perform about the same on both versions of the test. These results were similar to those in the normative sample for WMS-III and WMS-IV ($m = 10.20$, $SD = 3.00$ and $m = 10.5$, $SD = 3.20$, respectively), and most similar to the Major Depressive Disorder Adult clinical group ($n = 84$, ages 21-69; $m = 9.60$, $SD = 2.90$, $m = 9.90$, $SD = 3.2$).

These results suggest that changes in scaled score performance from VPA3 to VPA4 are less reliable as a means of measuring verbal explicit memory in clinical

outpatients for individual examinees than when examining group data. As a group, outpatient clinical neuropsychology examinees appear to perform similarly on VPA3 and VPA4. These data are consistent with, if slightly worse than, individuals from the WMS-IV normative sample, and similar to depressed individuals from the WMS-IV major depressive disorder clinical subgroup. At face value, group comparisons seem to support direct Time 1 to Time 2 comparisons from VPA3 to VPA4 in making clinical inferences about change over time in explicit memory functioning (which is the primary reason a neuropsychologist would administer both versions of the test).

However, a closer examination of current results suggest that a direct comparison may be contraindicated. While all participants performed within one standard deviation from VPA3 to VPA4, the degree of agreement was much less than hypothesized for VPA I and VPA II for WMS-III and WMS-IV. Possible explanations for these differences were offered above. Only about 50% of participants score within one scaled score point from one version of the test to the next. This finding is noteworthy because all participants completed both the WMS-III and WMS-IV as part of the same test battery, which would theoretically maximize the likelihood that persons would perform similarly, if only through practice effects. Additionally, it almost goes without saying that participants would not have sustained the type of brain dysfunction that would be expected to result in noteworthy score changes from one version to the next. Thus, despite well-controlled conditions designed to maximize internal validity and the likelihood of scaled score similarity, the lack of agreement was surprising. There are several practical implications as a result.

First, these results indicate that direct VPA3 to VPA4 comparisons may be

contraindicated in routine clinical practice. This study's relatively well-controlled conditions suggest that even under the best circumstances, individual examinees will perform within one scaled score point only about 50% of the time, at best. When real world confounds are introduced, such as the possibility of decline in functioning over time, these results suggest that changes in scaled scores from VPA3 to VPA4 cannot be attributed primarily to explicit memory performance.

Consequently, these results support the recommendation that clinical neuropsychology outpatient examinees presenting for serial assessment of memory functioning who have previously been assessed with the WMS-III should be reevaluated with the WMS-III rather than WMS-IV. This research supports this conclusion for WMS-III and WMS-IV Verbal Paired Associates. This issue was not addressed on other WMS-III/IV subtests or indices.

Second, these results indicate that practicing neuropsychologists should carefully consider the implications of using alternative norms when comparing tests that purport to measure identical constructs. This research used serial assessment within the same battery and found that while mean scaled scores are almost identical, changes in scaled scores on an individual level occur about half the time on a test that was hypothesized to agree more than 9 times out of 10.

Third, results of Hypothesis One suggests that comparisons from one test to the next should be examined with more scrutiny, using more rigorous research methodology than measures of central tendency such as mean, or measures of agreement across time using correlation coefficients. Specifically, this research indicates that a careful examination of intraindividual performance is indicated when assessing the extent to

which a novel version of a test measures the same construct as its predecessor. The methodology advanced here involved direct comparisons of degree of agreement derived from absolute difference scores from one version of the test to the next. More advanced methodologies using raw scores and inferential statistics would allow for the apriori development of disprovable hypotheses prior to beginning validation studies by test publishers. This could lead to more effective standardization studies by improving upon the common use of measures of central tendency across groups.

Fourth, the results of Hypothesis One indicate the continued need for additional implementation of neuropsychological, neuroanatomical, and neuropathological theory into the development of memory tests like Verbal Paired Associates. Per the WMS-IV Technical Manual, Verbal Paired Associates "measures the [...] ability to recall novel and semantically related word associations [...] low scores may indicate difficulty learning new associations" (Wechsler, 2009, p. 164). While technically accurate, this "interpretation" of performance leaves much to be desired. In practice, neuropsychologists are less concerned with whether an examinee learns and recall word associations as much as they are concerned about what the inability to learn or recall the association means in the context of known neuroanatomical and neuropathological dysfunction, such as is seen in Alzheimer's disease or major vascular neurocognitive disorder.

Fortunately, in the case of the WMS, subsequent research has demonstrated its validity in assessing for the patterns of memory impairment seen in numerous types of dementia, including those mentioned above. However, the results of this research suggest that reliance on group mean scores and correlational analyses alone at the exclusion of

more rigorous examination of intraindividual subtest score performance changes could be problematic. One would expect this issue to be most prominent in the months and years immediately after a new version of a test is published, before subsequent research can be conducted to guide decision-making about the appropriateness of substituting of version of the test for the next. Given that the WMS is updated about every 10 years, and the last version was published in 2009, the results of this research are both empirically relevant and timely.

Hypothesis Two

Hypothesis Two predicted that WMS-IV Verbal Paired Associates would show a significantly stronger relationship with attention, intellectual, and executive functioning ability as measured by performance on CPT-2 Commissions and Omissions, WAIS-IV, Trail Making Test Part B, and by the Category Test in a clinical sample than would WMS-III Verbal Paired Associates. This hypothesis was proposed because the revisions to the VPA subtest from WMS-III to WMS-IV (e.g., improved sampling methodologies, addition of semantically-related items) was expected to have a stronger relationship to intellectual, attentional, and executive functioning abilities in examinees referred for outpatient neuropsychological evaluation. Verbal Paired Associates is among the most widely-administered instruments used by neuropsychologists to assess explicit episodic memory performance. Hypothesis Two was not supported by the results, which indicate that VPA I and VPA II performance does not differ significantly from WMS-III to WMS-IV across WAIS-IV subtests and indices and measures of attention and executive functioning.

Most memory tests, including Verbal Paired Associates from the WMS, were

designed to be used by neuropsychologists for clinical purposes, such as cognitive impairments in clinical samples or deficits associated with aging, such as neurodegenerative diseases. The authors of the WMS-III changed VPA significantly from the WMS-R. VPA3 consisted of eight new word pairs that were semantically unrelated (WMS-R VPA consisted of four semantically related word pairs and four semantically unrelated pairs). The VPA3 word pairs are "hard" to learn in the sense they are semantically unrelated. The goal was to increase ecological validity by making VPA3 a purer test of associative learning than WMS-R VPA by removing semantically related items so that, in theory, all material learned reflected novel encoding, storage, and/or retrieval. Unfortunately, as described previously in reviewing the WMS-III factor analytic studies, what the test was supposed to measure (i.e., memory storage and retrieval after a delay) was not what it actually measured (i.e., auditory, visual, and working memory). In addition, VPA3 was criticized for the presence of ceiling effects for younger and healthier examinees and also for floor effects for less cognitively intact and older adults (Wechsler, 2009). Ceiling effects are common in most memory tests, including the CVLT-II (Delis, Kramer, Kaplan, & Ober, 2000) and the Rey Auditory-Verbal Learning Test (Rey, 1964). However, ceiling effects are not typically an area of concern for clinical purposes because memory testing is typically requested when deficits are suspected, and thus there has been an informal acceptance whereby specificity is sacrificed for increased sensitivity.

Floor effects are a problem however, when assessing impairment, however, and the authors of WMS-IV attempted to reduce them by creating an Older Adult battery that was shorter than the standard Adult battery, and for VPA4, four new semantically related

word pairs (i.e., “easy” word pairs) were added. For the Adult battery, each of the four trials consists of 14 items. For the Older Adult battery, each trial consists of 10 items. Practically speaking, the difference from VPA3 to VPA4 is that test should be “easier” because 29 to 40% of the items are semantically related, depending on which battery was administered (as compared to 0% on VPA3).

These results did not demonstrate a significant difference in scores from VPA3 to VPA4 for adult clinical outpatient neuropsychology examinees. There are several potential reasons why the changes to the test did not work as planned.

First, it is possible that the addition of semantically related word pairs has little to nothing to do with memory encoding, storage, or retrieval as assessed by paired associates tasks such as VPA in clinical participants. In healthy, cognitively intact persons, meaningful stimuli facilitates the processing of related stimuli or information. The semantic priming effects are understood to be a core component of how memory processes operate within a network model for long-term storage of information (Collins and Loftus, 1975). Recent research has shown that brain regions involved with semantic priming effects are less active in persons diagnosed with schizophrenia, with the net effect being that schizophrenic patients show no difference in brain activation regardless of whether items are semantically related. Further, these effects are thought to be correlated with severity of psychosis and the development and maintenance of delusions (Boyd, Patriciu, McKinnon & Kiang, 2014). These findings are relevant to the current research because they are based on EEG studies measuring reaction time (i.e., 400 ms post-stimulus onset in these studies) and are thus non-localizing to one specific brain area and perhaps to one type of mental disorder. One research implication of this research

then, is that it may be beneficial to assess for similar effects in other clinical populations.

A second reason why the changes from VPA3 to VPA4 did not may not have led to demonstratable differences for adult clinical neuropsychology outpatients is that the results may have been confounded by other changes made to the WMS-IV independent of VPA. Subjects in the WMS-III normative sample were inadequately assessed for cognitive dysfunction, which may have led to the inclusion of persons with impaired cognitive abilities, including memory. The presence of persons with mild cognitive impairment (mild neurocognitive disorder in DSM-5 parlance) in the WMS-III normative sample could have artificially lowered the mean performance of the sample, especially for those age groups most at risk for such conditions (i.e., older adults). WMS-IV sampling procedures included more advanced screening techniques for to exclude persons with suspected memory impairment. These procedural differences, when applied to the current clinical population, may have resulted in a cancelling out of performance effects, which would nullify any actual differences.

A third consideration for why VPA test changes may not have resulted in desired effects concerns the implementation of a separate test battery in WMS-IV for older adults. The Older Adult battery was developed to be shorter to mitigate the effects of performance fatigue for Older Adults. For VPA4 this led to a reduction in test items from 14 word pairs to 10. As discussed earlier, the number of semantically-related, “easy” items was held constant across batteries at four. In contrast, WMS-III did not have an adult battery, and the extent to which fatigue affected performance in the normative sample and the current clinical sample is unknown. Further, because participants in this study were selected from a convenience sample of consecutively seen outpatients, and

because the sample itself was relatively small, these factors may have contributed to a lack of appreciable differences in test performance from VPA3 to VPA4.

The current research indicates the need to consider several possible changes to improve the validity of VPA in future editions of the WMS. First, it is recommended that the test publisher consider and make explicit the theoretical rationale for decisions made concerning changes to VPA (or lack thereof). The changes to VPA from WMS-III to WMS-IV were made following psychometric examination of the normative sample and clinical samples (e.g., factor analytic studies, observed floor and ceiling effects), general complaints from neuropsychologists, patients, and third parties (e.g., potential negative performance effects due to fatigue; indirect pressure from managed care providers to assess cognitive abilities, including memory, more quickly, with fewer tests), and ongoing efforts from the test publisher to produce a product that reflects modern normative abilities for memory functioning. There is no evidence that changes made from VPA3 to VPA4 were informed by neuroanatomical, neuropathological, or any other empirical basis, despite the availability of such information dating to the 1950s. Moving forward, grounding changes in an empirical framework will allow for researchers and clinicians to evaluation the psychometric properties and ecological validity of VPA and the WMS from a stronger scientific position, which will ultimately serve consumers of the test (i.e., patients) better.

A second change recommended for VPA moving forward is increased attention to and transparency about the ecological validity of using paired associates as a means for assessing memory functioning in adults from a clinical population. As described earlier, there exists strong empirical evidence that some clinical populations fail to benefit from

semantic priming. This research should be evaluated critically and considered when deciding what changes to make to future editions of the test. Decisions such as altering the length of the test across batteries or to include semantically related word pairs should be informed by empirical evidence of ecological validity rather than the internal and external pressures of extraneous factors.

Finally, these results have important implications for neuropsychologists who are conducting serial assessments using WMS-III and WMS-IV. VPA3 to VPA4 scores in this sample did not reveal significant differences in performance for clinical outpatients. While on the surface this indicates that it could be appropriate to substitute one score for the other when making comparisons, a closer examination of the shortcomings of WMS standardization sample and the unclear effects on performance from changes to the structure of VPA from WMS-III to WMS-IV contraindicate the substitution of test scores when making diagnostic and prognostic decisions.

General Discussion

The current study sought to examine relationships between auditory episodic memory performance as assessed by VPA across two versions of the WMS (WMS-III and WMS-IV) and commonly assessed cognitive domains, including intellectual functioning, sustained attention, and executive functioning within an outpatient clinical neuropsychology sample. An important overall goal of the study was, to the extent results allowed, to inform neuropsychological research and practice through practical recommendations.

Experiences are transformed into memories through a series of complex

processes, including encoding, storage/recall, and recognition/retrieval. Assessment of memory functioning is one of the most common reasons adults are referred to neuropsychologists for a variety of reasons. Memory impairment is often a prominent sign and symptom for many acute (e.g., traumatic brain injury) and neurodegenerative (e.g., dementia) forms of neuropathology, and thus the identification of memory impairment is an important function that neuropsychologists provide. Memory impairments are also challenging for patients and their families, due to the critical role intact memory functioning plays in the management of basic and instrumental activities of daily living, occupational functioning, family relationships, and persons' individual identity. It is often impairment in these functional areas that lead patients and their families to seek out neuropsychological evaluations.

The first edition of the Wechsler Memory Scales was introduced nearly 80 years ago (Wechsler, 1945), though it has been in development since at least 1917 (Wechsler, 1917). Since its release, the WMS has been the instrument of choice for assessing memory impairment by neuropsychologists (Rabin, Barr, & Burton, 2005; Rabin et al., 2016). However, despite nearly eight decades of clinical use and three revisions to the WMS, memory assessment using the WMS (and other memory tests) continues to be plagued by a lack of theoretical grounding, and the technical manual provides very little information concerning test performance for clinical populations.

This research concerned the VPA subtests because the assessment of auditory episodic learning and memory is an integral component of most neuropsychological evaluations for clinical patients assessed on an outpatient basis and because the test was changed substantially from the WMS-III to the WMS-IV. These results of this research

suggest several new important findings. First, the degree of agreement between VPA3 and VPA4 is lower than expected for clinical outpatients. Participants in this study completed both WMS-III and WMS-IV, and it was predicted that 90+% of scaled scores would fall within one point from VPA3 to VPA4. Instead, only 41% fell within one point for VPA I and only 55% fell within one point for VPA II. These findings were noted in the context of overall similar subtest means across versions of VPA. These scores suggest there may be important performance differences for the VPA3 and VPA4. Possible explanations include 1) higher rates of cognitive dysfunction in the WMS-III normative sample combined with the addition of semantically related word pairs on VPA4 resulted in similar mean scores but not individual scaled scores and 2) the net effect of age (and possibly education) effects resulted in increased variance that is observable at the individual level but not when using mean scaled scores as a measure of central tendency.

Overall, the lack of agreement in scaled scores from VPA3 to VPA4 likely reflects the heterogeneity of clinical samples, and it serves as a reminder that applying nomothetic principles to idiographic situations can be problematic under even the most controlled circumstances, such as in the case with this research, where each participant served as their own control by completing each version of the test. The practical recommendation for neuropsychologists then, would be to take caution when comparing VPA3 to VPA4 results for individual patients. Under the best circumstances, clinical outpatients perform within one scaled score point only about half the time. Therefore, the effects of interim brain dysfunction via progressive decline or acute injury would be expected to be much less reliable across time. These results suggest that in situations such as these, it would be better to re-administer VPA3 rather than to administer VPA4 to

assess for changes in memory over time.

The present finding of discrepant scaled score agreement from VPA3 to VPA4 has also been observed in broader research with the WMS. As described earlier, factor analytic studies with WMS-III were notoriously discrepant, which is what led Hoelzle, Nelson, and Smith (2011) to recommend, "that WMS-III index scores be interpreted cautiously" (p. 290). Their investigation of WMS-IV was more promising: they were discovered a factor solution that, unlike WMS-III, adequately differentiated between auditory and visual memory performance for clinical patients. However, they emphasized the importance of heterogeneity of test performance with clinical populations: "There is conflicting evidence whether clinical and nonclinical samples should produce similar factor structures (2011, p. 290). Like the current research, the factor analytic studies in question were comprised of clinical samples, which inherently have more heterogeneity than nonclinical samples, such as the standardization samples upon which VPA3 and VPA4 scaled scores are obtained. Therefore, the current research indirectly supports the findings of Hoelzle and colleagues, and they directly support their conclusion that additional efforts to determine whether psychometric properties are consistent across distinct clinical sample is indicated to advance the field of neuropsychology by improving clinical assessment.

The current research, despite the shortcomings mention above, supports the continued use of VPA4 as a means of assessing the ability to form new associations. As a whole, clinical participants in this study performed about the same on VPA4 as they did on VPA3. Both versions measure retention of verbal paired associates, and both seem to measure the examinee's ability, on average, to retain that information following a 20 to

30-minute delay. If retention of information is synonymous with memory, which was David Wechsler's position, then VPA4 measures memory.

The current results are supported by other research investigating the usefulness of the paired associates modality with clinical populations, especially early Alzheimer's disease (AD; Blackwell et al., 2004; Fowler, Saling, Conway, Semple, & Louis, 2002; Lindeboom, Schmand, Tulner, Walstra, & Jonker, 2002). For example, Lowndes, Saling, Ames and colleagues, in their study comparing elderly patients with AD to healthy controls, found that "a verbal associate-recognition paradigm, containing arbitrarily associated words, can be as effective as a cued-recall analogue for discriminating patients in the early stages of AD from healthy elderly people" (2008, p. 595). Importantly, the authors found that the results were significant both at the group and individual level of analysis. An interesting caveat to their research was that they found that patients with AD performed poorly on concrete and abstract word pairs, which suggests that future versions of VPA might benefit from including all concrete words in the Older Adult battery.

With his introduction of a verbal paired associates task in his 1917 Method of Paired Associates, Wechsler found that patients diagnosed with Korsakoff's psychosis performed normally with semantically-related, "easy" word pairs (e.g., come-go, lead-pencil), but their performance was impaired for semantically-unrelated, "hard" word pairs. Since the first edition of the WMS easy and hard word pairs have been included in the VPA subtest (with the exception of WMS-III). The current findings suggest that more empirical data are needed to establish the usefulness of including both easy and hard word pairs in future versions of VPA. This recommendation is consistent with that of

other research that found no significant difference in easy versus hard word pairs in discriminating patients with mild amnestic cognitive impairment (aMCI), a known precursor to AD and other forms of dementia (Pike, Kinsella, Ong et al., 2013). The authors suggested the discrimination failure could be due to the fact that word pairs fail to tax the areas of the brain involved in aMCI and AD (i.e., the medial temporal lobe system). Collectively, the current research and prior research indicates the need for further investigation of the use of easy versus hard word pairs and reemphasizes the importance of grounding future versions of VPA in known biological and neuroanatomical mechanisms of memory functioning.

Limitations

There are several limitations to this study that potentially limit the widespread applicability of its results. One limitation of the study is the small sample size, which are associated with unintended consequences, including 1) lower statistical power, which may reduce the chances of finding true effects; 2) the production of results that have low reproducibility (Button, Ioannidis, Mokrysz et al., 2013). While statistical procedures were used to mitigate the effects (e.g., adjusted alpha levels to .01), the small sample size was certainly a limitation.

Another limitation of the current study involved the introduction of practice effects and/or interference as a result of administering both the WMS-III and WMS-IV to participants as part of the same test battery. There are several potential effects that could have detrimentally affected the study. First, practice effects may have primed participants' performance from VPA3 to VPA4, which could have resulted in better performance than if participants had only completed VPA4. While research has shown

that practice effects are present in healthy persons and those diagnosed with MCI for certain types of verbal episodic memory tasks (i.e., list learning), the extent to which practice effects carry over from VPA3 to VPA4 is unknown.

Second, for some participants, exposure to both versions of VPA could have introduced unintended interference into the learning process. Specifically, proactive interference effects may have caused reduced performance on VPA4 for some participants. Because data were not screened for interference effects on a case-by-case basis, one limitation of this research is that the extent to which potential interference effects resulted in performance changes is unknown.

Another potential limitation of this study involves the sampling procedure, namely, that the data were obtained from an archival dataset. As such, there was no way to screen for or control the sample characteristics (e.g., demographic factors, such as age, or psychiatric diagnosis). Further, there was no way to control for the order in which tests were administered. It is unknown whether all participants completed the WMS-III prior to the WMS-IV, for example. Further, it is unknown how much time elapsed between administration of VPA3 and VPA4. Given the length of the typical research battery within the Neuropsychology Assessment Center, it is reasonable to conclude that time between administration may have varied from as much as one or two days up to several months.

A second weakness related to the use of the archival dataset is the inherent lack of internal validity that accompanies the use of most research using archival data. Specific areas of concern for this research as they relate to internal validity include a lack of a control group (e.g., clinical participants with a relatively cognitively benign mental

disorder, such as adjustment disorder), lack of randomization (i.e., the study was essentially a convenience sample taken from consecutively seen patients – while this results in improved external validity, internal validity suffers), and a lack of pre-or post-tests (e.g., it is unknown whether these effects are stable over time or whether effects would vary over time within each participant as a result of age, mental health status, or other factors.

Another limitation of the current research involved the widespread age range of the participants, which introduced several problems for the research in terms of design and interpretation. First, the age range, which extended from young adults to the elderly, required the use of two different versions of the WMS-IV, one for adults aged through 66 to 69, and another for older adults above that cutoff. In contrast, all participants completed the same form for WMS-III. Specific to VPA, participants who completed VPA4 in the Adult Battery had to learn and recall 14 word pairs, while those who completed the Older Adult Battery were presented with only 10 word pairs to learn and recall. While it is hoped that the conversion of raw scores to scaled scores using the age-corrected normative procedures would account for this variance, the relatively small sample size of this research combined with the diversity of age and mental health status may have introduced unknown confounds when crossed with the two different test batteries on the WMS-IV.

Implications for Further Research

The empirical examination of neuropsychological tests from one version to the next has is critical for the field of neuropsychology. A thorough understanding of what our tests measure, and how that changes over time, has important implications for how

neuropsychology is practiced by clinicians (e.g., a newly released test may be less ideal from than its predecessor if changes to the test result in unintended and unwanted effects) and ultimately, for high-quality, ethical patient care. To continue moving toward this ambition, results of the current research have several implications for the future.

It is recommended that degree of agreement be reexamined when the next version of the WMS, WMS-5, is released. The test is currently in field trials until 2020 and will likely be released in 2021 or 2022. If consistent with prior versions of the WMS, it is expected that the test publisher will do an excellent job with the normative sample, and it is expected that the sample characteristics for healthy persons will be consistent with a broad spectrum of the U.S. population. However, this says little about how the test will measure memory in diverse clinical samples, and it is again expected that this issue will be left to subsequent researchers to investigate and publish on after the WMS-5 is released. A prospective study comparing how VPA from the WMS-5 assesses memory differs from prior versions of the WMS is needed to understand how the assessment of auditory episodic memory varies with clinical populations. Ideally, such a study would improve on the current research in several important ways.

First, an improved study would be prospective and allow for better control of factors that introduce variability and uncertainty in results and their interpretation. These include a large enough sample size to increase the power needed when assessing results over a variety of mental health diagnoses and age ranges. An outpatient convenience sample is not contraindicated, but the sample should be heterogenous enough in terms of age and mental health functioning to allow for generalization to the wider neuropsychological community.

Second, examination of raw and scaled scores is recommended to determine whether normative differences exist for memory functioning between nonclinical and outpatient neuropsychological individuals. A prospective study design would allow for just this type of in-depth analysis that may prove very relevant for the assessment of memory functioning moving forward. Third, researchers investigating the differences in memory performance between current and future versions of VPA and other memory tests would do well to search for potential difference in psychiatric inpatient participant populations, in addition to those hospitalized with comorbid medical conditions.

Another important implication of future research is that additional research is needed in the area of older adult performance within the mental health clinical populations. It is important to understand, for example, if elderly consumers of alcohol have important differences in memory performance than do non-drinkers. These differences may not simply be quantitative; rather, qualitative differences may also exist and warrant additional research with the publication of WMS-5. This issue will continue to increase in relevance as the U.S. population continues to age and many elderly persons present for neuropsychological evaluation with numerous comorbid medical and mental health concerns. This type of research could be accomplished by examining the raw score performance across tests and also by examining contrast scaled scores (introduced with WMS-IV) from one version of the WMS to the next. A look at process-oriented variables, such as is included in the expanded score report for the California Verbal Learning Test, Third Edition (CVLT-3) would also be helpful in understanding differences in memory functioning from one version of VPA to the next.

A final area of consideration concerns the longstanding need for the assessment of

memory to be grounded in sound theory of memory functioning as well as neuroanatomical models, of which have advanced remarkably over the past 50+ years since the WMS was originally released. For example, authors have commented on recent neuroanatomical findings in the specific ways in which discrete layers within the amygdalar region and hippocampal regions play a critical role in the learning and memory of emotional information through the mediation of GABA and glutamate projections. In this model, dysfunction has been described as a result of damage to one or more of these areas or to their interconnections, and the authors point out that both mental health disorders (e.g., PTSD) as well as organic disorders (e.g., AD) are implicated (McDonald & Mott, 2017).

Similarly, Gilpin and Weiner (2017), in an excellent review of the anatomical and biological models of comorbid PTSD and alcohol use disorder, described how persons with both conditions have important differences in both brain structure and function in terms of learning and memory. For example, they described how researchers using animal models found that acute exposure to alcohol facilitates the reactivation of existing memories from the past and dependency on alcohol leads to problems with subsequent extinction of fear. An important area of emerging memory research is the investigation of findings such as these in human participants.

In terms of autobiographical and sematic memory functioning, two popular theories exist and could be used for guidance during the development of memory tests of these systems. The first theory suggests that all declarative memory (semantic and episodic) become independent of the hippocampus as a function of time following learning through gradual changes to the neocortex. This theory, referred to the standard

consolidation theory, was well-described by Squire and Alvarez (1995) and provides substantial grounding for the development of memory tests, as described by (Kent, 2013).

The second theory involves the idea that the episodic content of a biographical memory is always dependent on the hippocampus, and that each time an episodic memory is retrieved, a copy of the memory is encoded into the hippocampus. Over time, more copies result in resistance to disruption in the memory. This theory, referred to as multiple trace theory or the transformation hypothesis, was advocated by Nadel and Moscovitch (1997) and elaborated and expanded by Winocur, Moscovitch, and Bontempi (2010).

These theories provide an empirical foundation for the development of models of memory functioning that could be applied to the development of memory tests. For example, they help understand why damage to the medical temporal lobe in isolation leads to anterograde amnesia (because it plays permanent role in the formation of new memories and the retrieval of autobiographical information). Moving forward, test developers are urged to use both these models of memory functioning and the recent advances in the neuroanatomical basis of memory to inform memory tests that are more grounded in empirically testable theories and less in quantitative analysis and modification.

REFERENCES

Axelrod, B. (2001). Administration duration for the Wechsler Adult Intelligence Scale-III and Wechsler Memory Scale-III (Vol. 16).

Axelrod, B., D. Dingell, J., Ryan, J., & L. Woodard, J. (2001). Cross Validation of Prediction Equations for Wechsler Memory Scale-Ill Indexes (Vol. 8).

Axelrod, B., Putnam, S. H., Woodard, J. L., & Adams, K. M. (1996). Cross-validation of predicted Wechsler Memory Scale-Revised scores (Vol. 8).

Basso, M. R., Harrington, K., Matson, M., & Lowery, N. (2000). Sex differences on the WMS-III: findings concerning verbal paired associates and faces. *The Clinical Neuropsychologist, 14(2),* 231-235. doi:10.1076/1385-4046(200005)14:2;1-z;ft231

Bornstein, R. A., & Chelune, G. J. (1988). Factor structure of the Wechsler Memory Scale-Revised. *The Clinical Neuropsychologist, 2(2)*, 107-115. doi:10.1080/13854048808520093

Bornstein, R. A., & Chelune, G. J. (1989). Factor structure of the Wechsler memory scale-revised in relation to age and educational level. *Archives of Clinical Neuropsychology, 4(1)*, 15-24. doi:https://doi.org/10.1016/0887-6177(89)90003-6

Burton, D. B., Mittenberg, W., & Burton, C. A. (1993). Confirmatory factor analysis of the Wechsler Memory Scale-Revised standardization sample. *Archives of Clinical Neuropsychology, 8(6)*, 467–475. doi:10.1093/arclin/8.6.467

Burton, D. B., Ryan, J. J., Axelrod, B. N., Schellenberger, T., & Richards, H. M. (2003).

A confirmatory factor analysis of the WMS-III in a clinical sample with cross validation in the standardization sample. *Archives of Clinical Neuropsychology, 18(6),* 629-641. doi:https://doi.org/10.1016/S0887-6177(02)00149-X

Cohen, W. (1950). Wechsler Memory Scale performance of psychoneurotic, organic, and schizophrenic groups. *Journal of Consulting Psychology, 14(5),* 371-375. doi:10.1037/h0062273

Comfrey, A. L., & Lee, H. B. (1992). A first course in factor analysis. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conners, C. K. (2000). Conners' Continuous Performance Test, 2nd edition. Toronto, Canada: Multi-Health Systems, Inc.

Corporation, T. P. (2002). WAIS-III WMS-III technical manual updated. San Antonio, TX: Author.

Davis Jr., L. J., & Swenson, W. M. (1970). Factor analysis of the Wechsler Memory Scale. *Journal of Consulting and Clinical Psychology, 35(3),* 430. doi:10.1016/0887-6177(86)90134-4

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). Manual for the California Verbal Learning Test, Second Edition (CVLT-II). San Antonio, TX: The Psychological Corporation.

Drozdick, L. W., Holdnack, J. A., & Hilsabeck, R. C. (2011). Essentials of WMS-IV assessment. Hoboken, NJ: John Wiley & Sons.

Dujovne , B. E., & Bernard, L. I. (1971). The psychometric structure of the Weschler

Memory Scale. *Journal of Clinical Psychology, 27(3),* 351-354. doi:10.1002/1097-4679(197107)27:3<351::AID-JCLP2270270313>3.0.CO;2-4

Elwood, R. W. (1991). The Wechsler Memory Scale—Revised: Psychometric characteristics and clinical application. *Neuropsychology Review, 2(2),* 179-201. doi:10.1007/BF01109053

Ernst, J., Warner, M. H., Morgan, A., Townes, B. D., Eiler, J., & Coppel, D. B. (1986). Factor analysis of the Wechsler memory scale: Is the associate learning subtest an unclear measure? *Archives of Clinical Neuropsychology, 1(4),* 309-314. doi:10.1093/arclin/1.4.309

Fields, F. R. (1971). Relative effects of brain damage on the Wechsler memory and intelligence quotients. *Diseases of the Nervous System, 32*, 673-675. Retrieved from https://psycnet.apa.org/record/1972-32023-001

Gass, C. S. (1995). A procedure for assessing storage and retrieval on the Wechsler Memory Scale-Revised. *Archives of Clinical Neuropsychology, 10(5),* 475-487. doi:https://doi.org/10.1016/0887-6177(95)98192-G

Golden, C. J., & Freshwater, S. M. (2002). The Stroop Color and Word Test: A manual for clinical and experimental uses. Chicago, IL: Stoelting.

Golden, C. J., White, L., Combs, T., Morgan, M., & McLane, D. (1999). WMS-R and MAS Correlations in a Neuropsychological Population. *Archives of Clinical Neuropsychology, 14(3), 265-271*. doi:https://doi.org/10.1016/S0887-6177(98)00017-1

Gilpin, N. W., & Weiner, J. L. (2017). Neurobiology of comorbid post-traumatic stress disorder and alcohol-use disorder. *Genes, Brain and Behavior, 16(1),* 15-43. doi:10.1111/gbb.12349

Haaland, K. Y., Linn, R. T., Hunt, W. C., & Goodwin, J. S. (1983). A normative study of Russell's variant of the Wechsler Memory Scale in a healthy elderly population. *Journal of Consulting and Clinical Psychology, 51(6),* 878-881. doi:10.1037/0022-006X.51.6.878

Hoelzle, J. B., Nelson, N. W., & Smith, C. A. (2011). Comparison of Wechsler Memory Scale-Fourth Edition (WMS-IV) and Third Edition (WMS-III) dimensional structures: improved ability to evaluate auditory and visual constructs. *Journal of Clinical and Experimental Neuropsychology, 33(3)*, 283-291. doi:10.1080/13803395.2010.511603

Hoffman, R., Tremont, G., G Scott, J., Adams, R., & Mittenberg, W. (1997). Cross-validation of predicted Wechsler Memory Scale-Revised scores in a normative sample of 25- to 34-year-old patients. *Archives of Clinical Neuropsychology, 12(7),* 677-682. doi:10.1016/S0887-6177(97)00022-X

Horton, J. A. M., & Larrabee, G. J. (1999). Wechsler Memory Scale III. *Archives of Clinical Neuropsychology, 14(5)*, 473-477. doi:10.1093/arclin/14.5.473

Howard, A. R. (1954). Further validation studies of the Wechsler memory scale. *Journal of Clinical Psychology, 10(2),* 164-167. doi:10.1002/1097-4679(195404)10:2<164::AID-JCLP2270100212>3.0.CO;2-M

Howard, A. R. (1966). A fifteen-year follow-up with the Wechsler Memory Scale.

Journal of Consulting Psychology, 30(2), 175-176. doi:10.1037/h0023182

Iverson, G. L. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology, 16(2),* 183-191. doi:10.1016/S0887-6177(00)00060-3

Kear-Colwell, J. J. (1973). The structure of the Wechsler Memory Scale and its relationship to 'brain damage'. *British Journal of Social and Clinical Psychology, 12(4),* 384-392. doi:10.1111/j.2044-8260.1973.tb00085.x

Kear-Colwell, J. J. (1977). The structure of the Wechsler Memory Scale: A replication. *Journal of Clinical Psychology, 33(2)*, 483-485. doi:10.1002/1097-4679(197704)33:2<483::AID-JCLP2270330233>3.0.CO;2-F

Kear-Colwell, J. J., & Heller, M. (1978). A normative study of the Wechsler Memory Scale. *Journal of Clinical Psychology, 34(2),* 437-442. doi:10.1002/1097-4679(197804)34:2<437::AID-JCLP2270340239>3.0.CO;2-K

Kent, P. L. (2016). Evolution of Wechsler's Memory Scales: Content and structural analysis. *Applied Neuropsychology: Adult, 24(3),* 232-251. doi:10.1080/23279095.2015.1135798

Kesner, R. (1973). A neural system analysis of memory storage and retrieval. *Psychological Bulletin, 80(3),* 177-203. doi:10.1037/h0034843

Larrabee, G. J., & Crook, T. H. (1995). Assessment of learning and memory. In R. L. Mapou & J. Spector (Eds.), Clinical neuropsychology: A cognitive approach (pp. 185-213). New York: Plenum Press.

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer program]. Retrieved from http://quantpsy.org

Loring, D. W. (1989). The Wechsler memory scale-revised, or the Wechsler memory scale-revisited? *Clinical Neuropsychologist, 3(1),* 59-69. doi:10.1080/13854048908404077

Loring, D. W., Lee, G. P., Martin, R. C., & Meador, K. J. (1989). Verbal and Visual Memory Index discrepancies from the Wechsler Memory Scale—Revised: Cautions in interpretation. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1(3),* 198-202. doi:10.1037/1040-3590.1.3.198

Luria , A. R. (1973). The working brain. New York: Basic Books.

Mahrou, M. L., Devaraju-Backhaus, S., Espe-Pfeifer, P., Dornheim, L., & Golden, C. J. (2000). Correlation of the WMS-III and measures of executive functioning. *Archives of Clinical Neuropsychology, 15(8),* 681-681. doi:10.1093/arclin/15.8.681

McDonald, A. J., & Mott, D. D. (2017). Functional neuroanatomy of amygdalohippocampal interconnections and their role in learning and memory. *Journal of Neuroscience Research, 95(3),* 797-820. doi:doi:10.1002/jnr.23709

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. Intelligence, 37(1), 1-10. doi:https://doi.org/10.1016/j.intell.2008.08.004

Mennemeier, M. S., Chatterjee, A., Watson, R. T., Wertman, E., Carter, L. P., & Heilman, K. M. (1994). Contributions of the parietal and frontal lobes to sustained attention and habituation. *Neuropsychologia, 32(6),* 703-716. doi:1016/0028-3932(94)90030-2

Migoya, J., Zimmerman, S., & Golden, C. J. (2000). Factor structure of the WMS-III in a neuropsychological population. *Archives of Clinical Neuropsychology, 15(8)*, 680-681. doi:10.1093/arclin/15.8.680

Millis, S. R., Malina, A. C., Bowers, D. A., & Ricker, J. H. (1999). Confirmatory factor analysis of the Wechsler Memory Scale-III. *Journal of Clinical and Experimental Neuropsychology, 21(1),* 87-93. doi:10.1076/jcen.21.1.87.937

Milner, B. (1968). Visual recognition and recall after right temporal lobe excision in man. *Neuropsychologia, 6(3),* 191-209. doi:10.1016/0028-3932(68)90019-5

Mittenberg, W., Burton, D. B., Darrow, E., & Thompson, G. B. (1992). Normative data for the Wechsler Memory Scale—Revised: 25- to 34-year-olds. *Psychological Assessment, 4(3),* 363-368. doi:10.1037/1040-3590.4.3.363

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion on Neurobiology, 7(2),* 217-227. doi:10.1016/S0959-4388(97)80010-4

Nott, P. N. (1975). The paired-associate learning subtest of the Wechsler Memory Scale: Six new parallel forms. *British Journal of Social and Clinical Psychology, 14(2),* 199-201. doi:10.1111/j.2044-8260.1975.tb00170.x

Powel, J. (1988). Wechsler memory scale-revised. *Archives of Clinical Neuropsychology, 3(4),* 397-403. doi:10.1093/arclin/3.4.397

Price, L. R., Tulsky, D., Millis, S., & Weiss, L. (2002). Redefining the Factor Structure of the Wechsler Memory Scale-III: Confirmatory Factor Analysis with Cross-Validation. *Journal of Clinical and Experimental Neuropsychology, 24(5)*, 574-585. doi:10.1076/jcen.24.5.574.1013

Prifitera, A., & Barley, W. D. (1985). Cautions in interpretation of comparisons between the WAIS-R and the Wechsler Memory Scale. *Journal of Consulting and Clinical Psychology, 53*(4), 564-565. doi:10.1037/0022-006X.53.4.564

Prigatano, G. P. (1978). Wechsler Memory Scale: A selective review of the literature. *Journal of Clinical Psychology, 34(4),* 816-832. doi:10.1002/1097-4679(197810)34:4<816::AID-JCLP2270340402>3.0.CO;2-Q

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(1),* 33-65. doi:10.1016/j.acn.2004.02.005

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in Test-Usage Practices of Clinical Neuropsychologists in the United States and Canada Over a 10-Year Period: A Follow-Up Survey of INS and NAN Members. *Archives of Clinical Neuropsychology, 31(3),* 206-230. doi:10.1093/arclin/acw007

Reitan, R. M., & Wolfson, D. (1985). The Halstead–Reitan Neuropsychological Test Battery: Therapy and clinical interpretation. Tucson, AZ: Neuropsychological

Press.

Roh, D. L., Conboy, T. J., Reeder, K. P., & Boll, T. J. (1990). Confirmatory factor analysis of the Wechsler Memory Scale-Revised in a sample of head-injured patients. *Journal of Clinical and Experimental Neuropsychology, 12(6),* 834-842. doi:10.1080/01688639008401025

Russell, E. W. (1975). A multiple scoring method for the assessment of complex memory functions. *Journal of Consulting and Clinical Psychology, 43(6),* 800-809. doi:10.1037/0022-006X.43.6.800

Russell, E. W., Neuringer, C., & Goldstein, G. (1970). Assessment of brain damage: A neuropsychological key approach. New York: Wiley-Interscience.

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion on Neurobiology, 5(2),* 169-177. doi:10.1016/0959-4388(95)80023-9

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87(2),* 245-251. doi:10.1037/0033-2909.87.2.245

Stone, C. P., & Wechsler, D. (undated). Wechsler Memory Scale manual. New York: Psychological Corporation.

Stuss, D. T., Alexander, M. P., Palumbo, C., Buckle, L., Sayer, L., & Pogue, J. (1994). Organizational Strategies of Patients with Unilateral or Bilateral Frontal Lobe Injury in Word List Learning Tasks. *Neuropsychology, 8*(3), 355-373. doi:10.1037/0894-4105.8.3.355

Tulsky, D. S., Chiaravalloti, N. D., Palmer, B. W., & Chelune, G. J. (2003). The Wechsler Memory Scale, Third Edition: A New Perspective. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik, R. A. Bornstein, A. Prifitera, & M. F. Ledbetter (Eds.), Clinical Interpretation of the WAIS-III and WMS-III (pp. 93-139). San Diego, CA: Academic Press.

Tulsky, D. S., Ivnik, R. J., Price, L. R., & Wilkins, C. (2003). Chapter 4 - Assessment of Cognitive Functioning with the WAIS-III and WMS-III: Development of a Six-Factor Model. In D. S. Tulsky, D. H. Saklofske, R. K. Heaton, R. Bornstein, M. F. Ledbetter, G. J. Chelune, R. J. Ivnik, & A. Prifitera (Eds.), Clinical Interpretation of the WAIS-III and WMS-III (pp. 147-179). San Diego: Academic Press.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psychology, 53*, 1-25. doi:10.1146/annurev.psych.53.100901.135114

Walton, D. (1958). The diagnostic and predictive accuracy of the Wechsler Memory Scale in psychiatric patients over 65. *Journal of Mental Science, 104(437)*, 1111-1118. doi:10.1192/bjp.104.437.1111

Wechsler, D. (1917). Retention defect in Korsakoff psychosis. *Psychiatric Bulletin of the New York State Hospitals, 2,* 403-451. Retrieved from https://scholar.google.com/scholar?cluster=11700387170460881164&hl=en&as_sdt=2005&sciodt=0,5

Wechsler, D. (1945). A standardized memory scale for clinical use. *The Journal of Psychology, 19,* 87-95. doi:10.1080/00223980.1945.9917223

Wechsler, D. (1955). The Wechsler Adult Intelligence Scale. New York: Psychological

Corporation.

Wechsler, D. (1981). Manual for the Wechsler Adult Intelligence Scale-Revised. New York: Psychological Corp.

Wechsler, D. (1997a). WAIS-III WMS-III technical manual. San Antonio: The Psychological Corporation.

Wechsler, D. (1997b). Wechsler Adult Intelligence Scale-III. New York: Psychological Corporation.

Wechsler, D. (1997c). Wechsler Memory Scale-Third Edition administration and scoring manual. San Antonio, Texas: The Psychological Corporation.

Wechsler, D. (2008). Wechsler Adult Intelligence Scale—Fourth Edition. San Antonio, TX: Pearson.

Wechsler, D. (2009a). Advanced Clinical Solutions for WAIS-IV and WMS-IV San Antonio, TX: Pearson.

Wechsler, D. (2009b). The Wechsler Memory Scale-Fourth Edition (WMS-IV). San Antonio, TX: Pearson Assessments.

Wechsler, D., Pearson Education, I., & PsychCorp. (2009). WMS-IV technical and interpretive manual. San Antonio, TX: Pearson.

Wechsler, D. A. (1987). Manual for the Wechsler Memory Scale-Revised. New York: The Psychological Corporation. Harcourt Brace Jovanovich.

Williams, J. M. (1991). Memory Assessment Scales professional manual. Odessa, FL: Psychological Assessment Resources.

Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: convergence towards a transformation account of hippocampal- neocortical interactions. *Neuropsychologia, 48(8),* 2339-2356. doi:10.1016/j.neuropsychologia.2010.04.016