UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DIGIT-SPAN RELATED PERFORMANCE VALIDITY INDICATORS IN PATIENTS WITH COGNITIVE IMPAIRMENT

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

MICHELLE HESTAND-OLSON

Norman, Oklahoma

2020

DIGIT-SPAN RELATED PERFORMANCE VALIDITY INDICATORS IN PATIENTS WITH COGNITIVE IMPAIRMENT

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Howard Crowson, Chair

Dr. John Linck, Co-Chair

Dr. Delini Fernando

Dr. Zermarie Deacon

Dr. Maeghan Hennessey

Acknowledgements

I would like to first and foremost thank my Internship supervisor, Dr. John Linck for introducing me to such an intriguing topic of research and providing seemingly unending support, knowledge, and guidance throughout my internship. I would also like to thank Dr. Mike Crowson, the chair of my committee and world’s best statistics professor. Without the guidance and support he provided, it is likely that I would still be trying to analyses and interpret the results of this study. Drs. Linck and Crowson have been with me through most of this process and I am eternally grateful. Special thanks to my other dissertation committee members Dr. Delini Fernando and Dr. Zermarie Deacon for being a positive presence with a supportive smile when needed. I would also like to thank our newest committee member, Dr. Maeghan Hennessey. I appreciate your willingness to step in and allow us to finish the journey.

To the faculty and staff of the Oklahoma University Health Science Center, I want to express my gratitude for allowing me to utilize the department’s archival data. I especially appreciate Mr. Gary Geiger for coordinating and assisting in accessing hundreds of charts involved in the data coding process.

To my family, you are the most amazing, supportive, loving, and sometimes maddening group of humans. Clay, you have been my partner and anchor in life. I wouldn’t want anyone else walking beside me on this crazy journey. Devon and Zachary, you have always been there to reminded me that the depth of my strength has no limit. It has been my honor to watch you grow into the amazing young men you have become. Being your mother has been my greatest joy. To my parents, you have faithfully encouraged, supported, and unconditionally loved me for my entire life. You taught me the true meaning of family. Everything I do in life, is in service to the incredible amount of love, support, and loyalty that my family has provided. Finally, I am grateful for all of the amazing supervisors that have been generous with their time and

knowledge. I will take the gifts that you provided and promise to never stop learning and growing as an individual as well as a professional.

Abstract

Neuropsychological evaluation is utilized to assess an individual's pattern of performance and level of cognitive functioning compared to a predicted premorbid level of functioning. Evaluation is accomplished through administration of neuropsychological battery of tests, following a diagnostic interview and review of medical records. Assessment of performance validity (i.e., effort or motivation) must be considered throughout the evaluative process. Performance validity (i.e., effort) is the concept that the obtained performance reflected in a patient's assessment profile is a true representation of that individual's ability, thus impacting the neuropsychologist's ability to interpret the obtained scores as being representative of their true cognitive functioning. While performance validity has been evaluated in many populations (Babikan, Boone, Lu & Arnold, 2006; Heinly, Greve, Bianchini, & Love, 2005; Greve et al., 2007), recent research has only briefly focused on the utility of effort indicators in the context of the performance of patients diagnosed with dementia (Kiewel, Wisdom, Bradshaw, Pastorek, & Strutt, 2012). The purpose of this retrospective review is to evaluate the efficacy of Digit Span-related performance as an indicator of poor effort in a clinical sample of patients diagnosed with either no, mild, or major neurocognitive impairment, ultimately identifying consistency between WAIS-IV Digit-Span related PVTs and if an individual's characteristics are significant predictor variables when determining performance validity. A logistic regression was utilized to analyze data to determine if the WAIS-IV Digit Span related PVTs are consistently measuring performance or if the determination of valid vs invalid responding is influenced by factors other than performance.

Keywords: neuropsychological evaluation, dementia, validity, performance

Table of Contents

Chapter 1:

Digit Span-Related Performance Validity Indicators in Patients with Cognitive Impairment

Neuropsychology is the study of the brain and behavior relationships (e.g., cognition, behavior, mood, and personality). The standardized processes neuropsychologists utilize to evaluate cognition is one of the best ways to identify possible deficits with neuropsychological testing demonstrating good sensitivity in identify neurocognitive dysfunction and distinguishing between dementias and non-dementias. Each evaluation will contain distinct processes and objectives, as well as answer specific questions. The process of evaluation includes standardized administration of a battery of neuropsychological tests (usually based upon the referral question), behavioral observations, conducting a diagnostic clinical interview (and collateral interview-when possible), a review of the patient's medical records, neuroimaging, patient questionnaires, and family/social history. The objective of a neuropsychological evaluation is three-fold: "1. Evaluate for the presence of a disturbance in higher cerebral brain functions; 2. Establish whether the pattern and level of test findings reveal diffuse, lateralized, or focal cerebral dysfunction; and 3. Does the resulting pattern and level of dysfunction correlate with a known or suspected organic disorder in a given patient?" (Prigatano and Pliskin, 2003, p. 15). Finally, the range of questions answered are broad, but generally fall under six categories: diagnoses, describing neuropsychological status, treatment planning, identifying the effects of treatment (measuring changes in functioning over time), research evaluation (impact of medication or cooccurring medical conditions), and forensic applications (Schoenberg & Scott, 2011). Although many providers administer cognitive tests, a clinical neuropsychological evaluation is distinguished from these more cursory reviews of neuropsychological function by the inclusion of a detailed, systematic assessment using psychometric tests with known standardized assessment procedures and normative performance data (Schoenberg and Scott, 2011).

When administering a battery of tests, it is important to be able to compare performance on tests that measure widely different skills. This comparison is made by an experienced Neuropsychologist trained in administration and interpretation of the test results. The easiest way to accomplish the task of comparison on performance, is to use standard scores rather than using raw scores. A raw score is a score that is presented in terms of the original test units. It is simply the number of items passed or points earned. A standard score, in contrast, is a derived score that uses, as its unit, the standard deviation of the population on which the developers standardized the test (Zillmer et al., 2008). The standard score is generally derived through transformation of a raw score using the mean and standard deviation associated with a particular population or norm group.

Standard scores allow for the establishment of norms, or expected scores based on similar demographic or educational backgrounds. Some commonly used factors when establishing normative data are age, gender, education, race and pre-morbid intelligence. Normative data sets are typically based on the normal curve (Zillmer et al., 2008) and are applied by comparing an individual’s test scores with the available normative data often obtained from individuals who are not suffering from the condition of interest (i.e. a patient suspected of having dementia will most likely be compared to others of a similar age and education level, who are not suffering from a known dementia). This approach provides the neuropsychologist with information regarding an individual’s ability in comparison with others. The method determines whether the obtained score of the patient in question deviates from the normative performance within the sample of interest and in what direction the deviation has occurred (i.e. does the patient performance the same, better, or worse than the normative group). The amount of quantitative deviation from the normative sample is then assigned a qualitative value (i.e. below average, average, above average) based on cut scores assigned to specific sections of the normal curve. A patient performing worse than the cut score may be labeled as having a weakness or impairment, whereas a patient scoring at or above cut score is labeled as intact with respect to that measure (Zillmer et al., 2008). It is that constellation of performances rather than a single measure that determines whether the individual is performing above or below expectations.

Most general approaches to neuropsychological interpretation of test performance, focus in some way on patterns or variability in performance by the individual. The pattern of performance is given assigned meanings, and certain patterns of performance are assumed to be reflective of specific neuropsychological syndromes when compared to known groups suffering from the condition of interest. Most often these interpretations are made solely on the basis of traditional norm referenced approaches (Reynolds, 1982). However, interpretation of these measures without adequate attention to the validity of the obtained data (i.e. attention to effort and motivation) could increase the risk for false positive and false negative errors.

The importance of objectively determining the validity of neuropsychological test scores has received a great deal of attention in the literature. Two major factors determine whether valid neuropsychological data will be obtained. First, the examiner must carefully adhere to all standardized administration and scoring procedures (Lee, Reynolds, & Willson, 2003, as cited in Miele, et al., 2012), which depends on the training of the practitioner and is difficult to objectively evaluate outside of a supervised test administration. The other factor is dependent on the examinee, whose degree of participation in the assessment determines the validity of the data. Suboptimal performance by an examinee for whatever reason invalidates the test findings

(Strauss, Sherman, & Spreen, 2006, as cited in Miele et al., 2012). It is absolutely essential, that all neuropsychological evaluations include methods of determining examinee effort throughout the assessment process (Miele, et al., 2012).

There is an emerging consensus among neuropsychologists that using a combination of multiple stand-alone and embedded PVTs should be routinely administered (Boone, 2009; Bush et al., 2005; Bush et al., 2014; Chafetz et al., 2015; Heilbronner et al., 2009, as cited in Erdodi & Abear, 2019). Many studies have examined performance validity in patients diagnosed with traumatic brain injury (TBI), in clinical research, and forensic cases (Heyanka et al, 2015). Much less is known about populations involving patients at various levels of cognitive impairment. In a study by Bortnick et al., 2013, the authors discuss some reasons why patients with neurodegenerative disorders may not have been as frequently researched as other populations, they found: “Many validation studies have excluded patients with dementia in part because of their generally lowered specificity rates and the fact that base rates of malingering are very low, with as few as 2% of litigants and those seeking other forms of compensation alleging vascular dementia (Mittenberg, Patton, Canyock, & Condit, 2002, as cited in Bortnick et al., 2013, p. 234). As a result, the efficacy of many symptom validity measures as they apply to dementia samples is largely unknown. Complicating matters further is the fact that if neuropsychological impairment is sufficiently severe, as in dementia, patients might fail effort measures despite putting forth adequate effort (Teichner & Wagner, 2007, as cited in Bortnick et al., 2013, p. 234).”

As described above, there is an assumption in the literature that due to the easy nature of performance validity tests, that individuals with neurologic problems can perform adequately on these measures (Heilbronner, et al., 2009), though much less is known about the ability to perform adequately on these measures across differing levels of impairment (e.g. mild, moderate, severe). This research is focused on identifying particular groups that may need the cut score to be adjusted in relation to their level of impairment. More specifically, most PVTs (embedded and stand-alone) are assumed to be appropriately easy for all individuals being administered the measure, however, the severity of an individual's impairment can greatly impact performance, yet the degree of severity is rarely mentioned with respect to the established cut scores for each measure. This research will attempt to address the discrepancies in test validity with populations of individuals 60+ and at varying degrees of impairment for Digit Span-related measures contained within the Wechsler Adult Intelligence Scale-Fourth Edition (Wechsler, 2008).

Literature Review

Overview of the Issue

Assessment of performance validity is an integral part of neuropsychological practice (Bortnik, Horner, & Bachman, 2013) occurring in the context of an ongoing evaluation. The development of Performance validity testing is largely a byproduct of forensic neuropsychology in which external incentives for suboptimal patient performance are often present, though poor effort can occur in non-litigating context as well. When observed, poor PVT performance raises the possibility that the entire neuropsychological assessment may be invalid (Loring et al., 2007), thus ensuring that a valid data set has been obtained is of extreme importance.

The neuropsychological evaluation produces data that needs to be organized and interpreted in order to determine level of impairment (if any), diagnosis, etiology, and treatment planning/recommendations in line with the processes and objectives described above. Evaluation is typically on an ordinal scale (i.e., impaired/non-impaired) for sensory/perceptual skills and progresses to an integral scale for more complex functions (i.e., assigned percentile obtained from converting a raw score to a standard score based on a distribution of scores from a normative group). It is generally known that any obtained score is a composite of both true score variance and error variance (Spreen, Sherman, and Strauss, 1991). Any factors (i.e. age, education, effort) that influence the true score can be considered in terms of error and reduce the level of confidence in the obtained score. While factors like age and education can be accounted and controlled for in advance, factors like effort are less controllable and more difficult to predict. The nature and design of assessment measures are especially susceptible to any attempt by the patient to malinger (Ziglar & Boone, 2015), because it is largely based on the patient’s performance. It is impossible to accurately interpret the data without reflecting on whether the patient is adequately engaged in the testing process. Simply asking if someone is trying hard or basing the assessment of effort solely on the appearance of giving full effort are problematic at best. Before addressing the types of PVT measures available and delving into the research on PVT performance in various dementia groups, an applied clinical example that highlights the author’s concern appears prudent at this stage. Consider the following two examples:

> Patient 1 is a 55-year-old, male with 16 years of education with no family history of dementia, presenting with an 18-month history of cognitive decline first observed by his co-workers (normal neuroimaging). The patient himself has not observed any changes and given he lives alone with no immediate family members in the area, a collateral visit was not possible. Assessing his Instrumental Activities of Daily Living (IADLs) is made more challenging due to the lack of a collateral informant, though he performs poorly on a functional measure assessing skills for check writing, counting change, and managing medications. The patient would like to continue working, has no history of legal

> problems, no mood disorders, no substance use history and is largely unconcerned about his thinking. During the evaluation, he is observed to perform poorly on standard cognitive measures but does not seem to notice. Complicating the clinical picture, he fails 2/4 EPVTs and has a "marginal" performance on a SPVT.
>
> Patient 2 is also a 55-year-old, male with 16 years of education with no family history of dementia, presenting with an 18-month history of cognitive decline following a concussion in the work setting. He has filed a disability claim, stating that he has cognitive deficits associated with his concussion (normal neuroimaging). He does not have a history of substance use or a history of mood disorders prior to his injury, but reports concerns about "post-traumatic stress" during his visit. He reportedly continues to perform his Instrumental Activities of Daily Living (IADLs) without difficulty per his wife's report, though assessment of his IADLs on a functional measure reveals problems with check writing, counting change and managing medications. He also performs poorly on a number of cognitive measures. Like Patient 1, he demonstrates a "marginal" SPVT performance and failed performances on 2/4 EPVTs.

Are either of these patient's putting forth sufficient effort? What are the implications of concluding that Patient 1 is putting forth insufficient effort if he is indeed impaired (i.e. false positive error)? What are the implications of concluding that Patient 2 is indeed impaired if this is not the case (i.e. false negative error)? These questions are answered through a comprehensive approach that takes into account factors such as knowledge and expectations regarding the condition of interest and consistency of self-report and observed performance, but also psychometric factors such as base-rates and resulting sensitivity and specificity of the PVT measures used in the assessment.

As noted above, PVTs assess effort through administration of an easy task that can typically be passed by those with neurologic, psychiatric, or developmental problems (Heilbronner et al., 2009). Essentially, tasks are developed to assess skills such as repeating digits forwards after they have been read or recognizing pictures immediately after they have been presented, which are skill that are typically preserved even in the face of neurologic injury or disease. Patients do not know in advance which measures are PVTs and which are not, as advanced knowledge would spoil the test.

It should be understood, poor effort does not equate malingering. The American Academy of Clinical neuropsychologists (AACN) Consensus Conference Statement on the Neuropsychological Assessment of Effort, Response Bias, & Malingering, highlights the dimension of effort levels existing on a continuum; these levels can fluctuate throughout an evaluation. Effort can be influenced by malingering, but can also be influenced by somatization, conversion, factitious disorder, or various other sources of poor motivation and opposition, which are not consistent with malingering (Strauss, Sherman, & Spreen, 2006). While intentional deceit is one consideration when interpreting poor effort scores, factors related to interest in the evaluation, psychiatric issues, substance use, or a lack of understanding of the assessment process should also be considered. However, a fundamental understanding of the issue of malingering is important.

Criteria for Malingering

There has been a widespread and concerted research focused on efficient methods for detecting exaggeration or fabrication of cognitive dysfunction. Despite these psychometric advances, the process of diagnosing malingering remains difficult and largely idiosyncratic (Slick, et al., 1999). Malingering of Neurocognitive Dysfunction (MND) is the volitional exaggeration or fabrication of cognitive dysfunction for the purpose of obtaining substantial material gain or avoiding or escaping formal duty or responsibility. Substantial material gain includes money, goods, or services of nontrivial value (e.g., financial compensation for personal injury). Formal duties are actions that people are legally obligated to perform (e.g., prison, military, or public service, or child support payments or other financial obligations). Formal responsibilities are those that involve accountability or liability in legal proceedings (e.g., competency to stand trial) (Boone, 2007).

Over the years there have been efforts to define poor effort and malingering by research groups and via formal diagnostic manuals. Nies and Sweet (1994), defined guidelines eventually utilized in the Slick Criteria for the development of the proposed malingering criteria, namely, the need for: (1) a specific definition of malingering of cognitive dysfunction within the context of the neuropsychological assessment; (2) specific, unambiguous, and reliable criteria that cover all possible sources of evidence (i.e., test-performance, observations, and collateral data); (3) specification of the relative importance of diagnostic criteria; (4) specification of the nature and role of clinical judgment; (5) specification of differential diagnoses and exclusionary criteria; and (6) specification of levels of diagnostic certainty.

The Diagnostic and Statistical Manuel of Mental Disorders- Fifth Edition (DSM-5; APA, 2013) defines malingering as “the intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives (e.g. avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs)” (DSM-5, 2013, p.726-727). Under some circumstances, malingering may represent adaptive behavior (e.g. feigning illness while a captive of the enemy during wartime). Malingering should be strongly suspected if any combination of the following is noted: (1) Medicolegal context of presentation (e.g., the individual is referred by an attorney to the clinician for examination, or the individual self-refers while litigation or criminal charges are pending). (2) Marked discrepancy between the individual's claimed stress or disability and the objective findings and observations. (3) Lack of cooperation during the diagnostic evaluation and in complying with the prescribed treatment regime. (4) The presence of antisocial personality disorder. Notably, malingering is not considered a mental health disorder. Additionally, it differs from factitious disorder in that the motivation for the symptom production in malingering is an external incentive, whereas in factitious disorder, the incentive is likely psychological in nature. Malingering is differentiated from Conversion disorder and somatic symptoms related mental disorders by the intentional production of symptoms and by the obvious external incentives associated with it.

Slick, et al. (1999) proposed set of diagnostic criteria that define psychometric, behavioral, and collateral data indicative of possible, probable, and definite malingering of cognitive dysfunction, for use in clinical practice and for defining populations for clinical research. They defined malingering as: "Malingering of Neurocognitive Dysfunction (MND) is the volitional exaggeration or fabrication of cognitive dysfunction for the purpose of obtaining substantial material gain or avoiding or escaping formal duty or responsibility. Substantial material gain includes money, goods, or services of nontrivial value (e.g., financial compensation for personal injury)" (Slick et al., 1999, p. 552).

According to the Slick Criteria, malingering was distinguished from potentially similar appearing presentations that are not part of a volitional attempt to obtain readily identifiable and commonly accepted external incentives. Examples of such presentations include poor or inconsistent effort, as well as defensive, hostile, or oppositional approaches to test taking that result from fatigue, psychiatric disturbance, and legitimate neurological impairment. define psychometric, behavioral, and collateral data indicative of possible, probable, and definite malingering of cognitive dysfunction, for use in clinical practice and for defining populations for clinical research (Slick, et al.,1999). In summary, the Slick criteria was designed to detect and define malingering as the fabrication, feigning, or exaggeration of physical or psychological symptoms designed to achieve a desired outcome.

Stand-alone Validity Measures

Traditionally, PVTs were stand-alone or free-standing instruments designed exclusively to monitor performance validity. Essentially, these are tests that “stand-alone” in term of only evaluating performance validity and are added to an existing battery of tests in order to ensure the test data is valid via adequate effort. Although they have robust classification accuracy (Larrabee, 2012), they are not without limitations. First, stand-alone PVTs use valuable resources as they are added to an existing battery and are not inexpensive to purchase and administer (e.g., clinician time and test materials). Second, they only provide data on the credibility of the response set, without informing diagnostic considerations, which is the main goal of the evaluation. Third, PVTs involving multiple trials/time delays place restrictions on the administration sequence of ability tests (Ryan et al., 2010). Lastly, PVTs only sample performance validity at discrete points in time (Erdodi & Abeare, 2019).

The Test of Memory Malingering (TOMM) and the Word Memory Test (WMT) are two examples stand-alone PVTs designed to be relatively impervious to central nervous system dysfunction (Green, Flaro, & Courtney, 2009) and are effective in identifying suboptimal effort with varying degrees of sensitivity and specificity (Tombaugh, 1997, p. 263). Additionally, the TOMM possesses a high degree of specificity and is not affected by demographic variables such as age and education (Rees et al., 1998).

The Test of Memory Malingering (TOMM) is one of the oldest, most widely used, and well-validated PVTs (Donders, 2005; Rees, Tombaugh, Gansler, & Moczynski, 1998; Sollman & Berry, 2011, as cited in Kraemer et al., 2020). The full TOMM consists of two learning trials (TOMM-1 and TOMM-2) and an optional retention trial (TOMM-R). TOMM-1 and 2 are administered by presenting 50 pictures of common objects with a 3-second presentation time followed by a 1-second interval between presentations. Upon completion of both learning phases, 50 recognition panels with two choices each are presented individually. Its primary purpose is to detect malingering of memory impairments, which requires a considerable time investment for administration (i.e., approximately 15–20 min; Kraemer et al., 2020).

Another widely used and researched PVT of cognitive performance is the computerized Green's Word Memory Test (WMT; Green, 2003 as cited in Donders and Strong, 2013). This instrument has adequate sensitivity and high specificity with regard to the detection of atypical effort in a variety of neurological conditions (Greve, Curtis, & Bianchini, 2013). Administration occurs via the presentation of 20- word pairs over two learning trials. Immediately following the learning trials, an immediate recall trial requires the examinee to choose between word pairs containing a word from the list and a non-list work. Following a 30-minute delay, Multiple Choice, Paired Associates, and Free Recall subtests are administered, in that order, which is then followed by an additional 10-minute delay and another long-delay free recall task. During the multiple-choice subtest, the first word from each of the original word pairs is shown on the computer screen, and the examinee has to choose the correct second word from eight options that are also shown on the screen. On the Paired Associates subtest, the examiner presents orally the first word from each original pair, and the examinee is then asked to provide the second word (without access to the computer screen). On Free Recall and Long Delayed Free Recall, the examinee is asked to recall as many words as possible from the original list of word pairs (again, without access to the computer screen; Donders et al., 2013).

The Victoria Symptom Validity Test (VSVT) is another commonly utilized stand-alone measure. The VSVT is a forced-choice PVT consisting of 48 five-digit stimuli presented in series of 16 stimuli with recognition delays of 5, 10, or 15 seconds. Each stimulus is presented for 5 seconds on a computer screen. After the brief delay, the target stimulus is presented with a foil and the subject indicates which of the two stimuli is the target. Half of the stimuli are easy targets in which foils sharing no common digits with the target are used; hard targets are contrasted with foils in which two of the digits have been transposed (Loring et al., 2007).

There are some effort measures such as the Word Memory Test (WMT; Green et al., 1996, as cited in Bortnik, et al., 2013), Medical Symptom Validity Test (MSVT; Green, 2004, as cited in Bortnik, et al., 2013), and the Nonverbal Medical Symptom Validity Test (NV-MSVT; Green, 2008, as cited in Bortnik, et al., 2013) that are designed to differentiate between suspect effort and severely impaired cognition through a ‘‘dementia profile,’’ with reported specificity levels of 89 percent to 98.5 percent in patients with dementia (Henry, Merten, Wolf, & Harth, 2010; Howe, Anderson, Kaufman, Sachsa, & Loring, 2007, as cited in Bortnik, 2013).

Because the nature of this project primarily relates to the following section on embedded validity measures and their use in the dementia population, the author has elected to not delve into specific studies associated with the measures described above aside from noting that they are robust in terms of their psychometrics and their classification accuracy. Conversley, as will be described below, it is less common for these stand-alone measures to be utilized in the dementia setting for a number of reasons associated with time, resources management, and fatigue effects in what are often shortened evaluation periods.

Embedded Validity Measures

In contrast, EPVTs are “after-market” cutoffs added to existing cognitive tests. More plainly, they measure cognitive ability and performance validity simultaneously without requiring additional administration time or test material. The use of an EPVT is attractive since they do not increase overall testing time. Additionally, they are cost-effective and are resistant to the effect of coaching and reduce the appearance of clinician bias toward malingering detection (Boone, 2013). In addition, there are some instances where embedded PVTs are the only tool available to the neuropsychological practitioner for assessing an examinee’s level of effort. Lastly, they provide continuous monitoring of performance validity throughout the test battery. Again, as previously noted, a measure that is considered stand-alone (TOMM, WMT, VSVT) is specifically designed to only measure the effort of the individual completing the test (Boone, 2013).

Selecting the appropriate type of validity measure (e.g., embedded or stand-alone) will vary as a function of the referral question in addition to time constraints, level of patient fatigue, level of cognitive impairment, medical conditions, and other variables that impact test selection. There are a number of embedded PVTs to choose from that assess effort across a range of cognitive domains given that poor effort is not specific to memory dysfunction alone. Examinees may simulate various types of impairment ranging from language dysfunction, problems with planning and problem solving, to motor dysfunction. As with stand-alone measures, the sensitivity and specificity of embedded validity indices to suboptimal effort varies.

One of the most commonly used EPVT is the Rey Auditory Verbal Learning Test (RAVLT), it is a neuropsychological assessment designed to evaluate verbal memory in patients, 16 years of age and older. The RAVLT consists of a list of 15 unrelated words repeated over five different trials, where patients are asked to repeat the list after each reading. A score is recorded after each of the five trials and combined to make-up the Total score at the end of the five immediate recall trials. Another list of 15 unrelated words (Distractor List) are then read and the patient must again repeat the original list of 15 words then and again after 30 minutes. Approximately 10 to 15 minutes is required for the administration (not including 30 min. interval; Strauss et al., 2006). The RAVLT can be used to evaluate the nature and severity of memory dysfunction and to track changes in memory function over time. The RAVLT EPVT, is calculated using the total learning raw score (Total) and the delayed recognition raw score (Rec).

Following the development of the measure, Davis, et al. (2012), calculated the sensitivity and specificity of the new index using a heterogeneous sample of 130 patients with mild traumatic brain injury (mTBI). The sample did not include patients with suspected dementia, neurological or psychiatric conditions (Davis, et al., 2012). After examining the data, they proposed two cutoff scores. The first, at the 50th percentile, demonstrated specificity of 81 percent and sensitivity of 68 percent. The second and more conservative cutoff, at 71 percent, demonstrated specificity of 91 percentile and sensitivity of 55 percentile (Poreh, et al., 2017).

Since the RAVLT is one of the most widely used neuropsychological tests in the literature and is applicable to a wide range of clinical groups, there is value in comparing performance validity across various groups from archival and previously published data (Malloy-Diniz, Lasmar, Gazinelli, Fuentes, & Salgado, 2007; Poreh, Sultan, & Levin, 2012; Schoenberg et al., 2006).

In a study examining some limitations of The RAVLT EPVT for detecting performance validity, the performance of four groups of 879 participants comprised of 464 clinically referred patients with suspected dementia, 91 forensic patients identified as not exhibiting adequate effort on other measures of response bias, 25 patients with well documented TBI, and a random sample of 198 adults collected in the Gulf State of Oman. The measure was also put to the test using normative data collected from the literature. Using sensitivity and specificity analyses, the results indicate moderate to high sensitivity yet low specificity. Using multisampling archival data, the study shows that utilizing the RAVLT as a EPVT was able to generate reasonably low false positives and moderate false negatives when it is employed in forensic practice, assuming the patients all have sustained mild TBI. Similarly, older adults with intact scores on a mental status exam, patients with well localized TBI, and normal controls can also be identified as exhibiting good effort. However, the measure was unable to properly distinguish between forensic patients who exhibit noncredible responses and patients with dementia (Davis, et al., 2012). Similar results were obtained when the measure was calculated using normative clinical data published in the literature. Specifically, in most studies of simulators, the index did not accurately detect a large proportion of the subjects as exhibiting noncredible performance, implying that they are suffering from a genuine neurocognitive impairment. The findings of the current study, using archival data as well as data published in the literature, also produced variable findings with regard to the sensitivity of the new embedded index. Specifically, they show that the specificity of the new index is low. Namely, many of the subjects who exhibited noncredible performance were still not identified, and at the same time, some of those who were identified as exhibiting noncredible performance were likely to be genuine cases (Poreh, Tolfo, Krivenko, & Teaford, 2017). This study is important to this research because it highlights the gap in research regarding sensitivity and specificity of the RAVLT as an EPVT in patients at various levels of cognitive impairment. Additional research is needed to differentiate between true positives and false positive effort scores.

Another assessment with a commonly utilized EPVT that is the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Paulson, et al., 2015). The RBANS was initially published in 1998 and developed as an assessment tool for dementia. The RBANS was designed to measure functional limitations in patients with dementia and mild cognitive impairments, used to be a brief neuropsychological evaluation, and used in longitudinal research (Badenes, Casas, Cejudo, & Aguilar, 2008; Duff et al., 2010). The structure of the RBANS is made up of five cognitive domains: Five Cognitive Domains: immediate memory, visuospatial/constructional, language, attention, and delayed memory (made up of 12 subtests). It includes four parallel forms designed to reduce the test-retest effects. It was originally normed for age ranges starting at 20 through 89 but was modified to cover 12 through 89:11 when RBANS Update was released in 2012. Administration of the RBANS usually takes approximately 20- 30 minutes and can be administered via Digital or Paper format. Scores are interpreted by calculating index standard scores, sub-test standard scores, and a total score. It can be scored manually or through a web-based program. RBANS is available in over 20 languages, which makes it one of the more versatile EPVTs (Strauss, Sherman, & Spree, 2006).

Additionally, the diagnostic accuracy of the RBANS has been shown to adequately detect cognitive impairment associated with Alzheimer’s disease (AD; Duff, Humphreys, Clark, et al., 2008). Although several studies have used the RBANS as a tool to examine cognitive dysfunction, there remains little information regarding the diagnostic accuracy of the RBANS and its ability to detect milder deficits in cognition in the elderly (Duff et al., 2010).

According to Paulson, et al. (2015), Novitski, Steele, Karantzoulis, and Randolph (2012), examined the RBANS Effort Scale (ES), an EPVT. As well as the RBANS Effort Index (EI). The RBANS ES attempts to identify invalid responding on the basis of large disparities between recall and recognition. This scale is applied in two steps. First, respondents with combined raw scores on the RBANS digit span and list recognition subtests greater than or equal to 28, whose responses are likely to be valid representations of their cognitive functioning based on data from the normative sample, are excluded. In the second step, the RBANS ES formula is applied and invalid responding is identified as scores equal to 12 (Novitski, et al., 2012, as cited in Paulson, 2013). Preliminary research using a sample of patients with dementia, and "coached and naïve" simulators supports the use of the RBANS ES, with those populations (Paulson et al., 2012).

The RBANS ES includes multiple subtests (e.g., List Recognition, List Recall, Story Recall, Figure Recall, and Digit Span). It identifies invalid responding based largely on the discrepancy between recall and recognition, another approach to delineating embedded measures draws on work identifying consistently low scores across disparate domains of cognition as a correlate of invalid performance validity (Meyers, Volbrecht, Axelrod, & Reinsch-Boothby, 2011; Schutte, Millis, Axelrod, & VanDyke, 2011).

According to Silverberg, et al., 2007 the RBANS EI seeks to identify invalid responding based on low scores on both digit span and list recognition. Scores on these two subtests are converted using a weighting algorithm based on normative score distributions. Converted subtest scores are then added to calculate the RBANS EI score. Their preliminary findings suggested good sensitivity (86–96 percent) and specificity (78–96 percent) in a mixed sample of individuals with mild traumatic brain injury, clinical malingerers, and coached and uncoached-simulated malingerers. Subsequent work found that this scale offered only modest predictive utility with specificity of 85 percent, and sensitivity ranging from 51 to 64 percent, based on varying cutoff scores (Barker et al., 2010, as cited in Paulson, et al., 2015). Similarly, Hook, Marquine, and Hoelzle (2009), reported that the RBANS EI may offer limited utility, particularly with geriatric medical patients (Paulson, Horner, & Bachman, 2015).

The majority of research literature on EPVTs include various editions of the Wechsler's Adult Intelligence Scales (WAIS), focusing on either a single subtest or a number of indicators nested within the same domain (i.e., working memory or processing speed). Since the test was originally conceived as a fixed battery of tests designed to produce a global measure of intellectual functioning, this study was designed to replicate that model for the emerging function of the WAIS-IV (i.e., performance validity indicator). Based on previous research demonstrating the superiority of multivariate assessment models in general (Meyers et al., 2014; Pearson, 2009; Tyson et al., 2018) and to minimize false-positive errors, specifically (Larrabee, 2008, 2014; Odland et al., 2015, as cited in Erdodi & Abear, 2019), combining the existing EPVTs into an aggregate validity index (i.e., a EPVT analog of a Full-Scale IQ) would improve classification accuracy of psychometrically defined non-credible responding as compared to univariate cutoffs (Erdodi & Abear, 2019).

The Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) is a widely used measure of cognitive functioning (Wechsler, 2008) and the host of several EPVTs. In a mixed clinical sample of adults referred for neuropsychological evaluation in a medical setting, every one of the eight core subtests of the WAIS-IV were significant predictors of psychometrically defined invalid performance (Erdodi & Lichtenstein, 2017). The repurposing of WAIS-IV subtests as EPVT started in the early 1990s, with the discovery of critical thresholds on Digit Span (DS) and Coding, which credible impairment was rare (Trueblood & Schmidt, 1993;

Trueblood, 1994). Subsequent research confirmed the utility of these EPVTs (Axelrod, et al., 2006; Etherton, et al., 2006; Heinly, et al., 2005; Kim, et al., 2010; Spencer, et al., 2013, as cited in Erdodi & Abear, 2019) and introduced validity cutoffs for the Symbol Search (SS), the Processing Speed Index (PSI) (Curtis et al., 2009; Erdodi, Abeare, et al., 2017; Inman & Berry, 2002, as cited in Erdodi & Abear, 2019), and Letter–Number Sequencing (LNS) (Shura, et al., 2016, as cited in Erdodi & Abear, 2019).

Further research confirmed that when examining predictors of suboptimal performance, the Reliable Digit Span (RDS) was a significantly better predictor of suboptimal effort than the other EPVTs (Poreh et al., 2017). RDS is calculated by adding the longest digit span correctly responded on both trials on the forward and backward subtests. An RDS of less than or equal to 7 was able to correctly classify in 74 percent of examinees. This classification rate is consistent with that reported in a recent meta-analysis that found a classification accuracy rate of 76 percent using RDS at the same cutoff (Jasinski, Berry, Shander, & Clark, 2011). The use of RDS alone provided a more parsimonious method for identifying suboptimal effort; examinees who failed RDS were almost eight times more likely to have failed two or more SVTs. Additionally, this study calculated the sensitivity and specificity for each embedded validity index. As reflected in the sensitivity and specificity values, each was associated with an elevated false-positive and false-negative rate (Poreh et al., 2017).

Additionally, the study by Poreh et al., (2017) demonstrated the lack of consideration that has been given in the past to reducing the frequency of false negatives and false positives. After reviewing both the inferential and descriptive data, Poreh et al., (2017), found that RDS emerged as the “most robust” EPVT in classifying suboptimal effort as defined by failing two or more stand-alone PVTs , although there was nearly a 20 percent chance of over-identifying suboptimal effort (false positives) and a 40 percent chance of missing suboptimal effort (false negatives), when applying the RDS only. Thus, there is not any one assessment that should be given enough weight to dictate performance validity or support any specific diagnosis. It is important to interpret these and any findings in the context of the overall evaluation.

An example of research supporting the combination of embedded and free-standing validity measures is a study by Paulson, et al. (2017). They found that the strategy of combining embedded and free-standing tests is consistent with Boone (2009) and more recently Larrabee's (2015) notions regarding the continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. The EPVT is an important and useful index for assessing response bias when it is administered in conjunction with dedicated measures as it allows clinicians to go beyond general statements regarding the validity of the test protocol and make informative statements regarding the validity of the patient's test scores (Poreh et al., 2017).

Dementia and Embedded PVT Performance

Although many studies have demonstrated the efficacy of EPVT in a variety of medical and compensation-seeking contexts, much less is known about the robustness of these measures in elderly populations, particularly in patients with dementia. Kiewel, et al. (2012), postulated, that there has been extensive research on the use of both stand-alone and embedded measures of effort in neuropsychological testing, though relatively few studies have reported on their utility in the context of genuine cognitive impairment. Previous studies that have examined the specificity of traditionally used cut-scores on embedded measures of effort with dementia samples have largely found high rates of false positive errors (Kiewel, et al., 2012). Although older adults may be viewed as less likely to intentionally feign symptoms for an external gain, there are a variety of other factors that could result in suboptimal effort, including fatigue, lack of interest or cooperation in the testing process, or failure to fully appreciate the implications of the assessment on treatment care and outcome (Bortnik, et al., 2013).

Much less research has focused on the predictive validity of EPVTs associate with specificity. Given the potential ramifications for the patient of misattributing genuine impairment as being the result of ‘‘suboptimal effort’’ (e.g., denial of a disability claim) it is generally recommended that the false positive error rate be less than 10 percent for any SPVT or EPVT (Larrabee, 2008), thus recommended specificity levels are set at .90.

Given the lack of research on cognitively impaired groups suffering from neurodegeneration, it is possible that specificity for poor effort could be lower in those with greater levels of cognitive impairment (Merton, Bossink, & Schmand, 2007). One way that researcher have dealt with validity concerns in populations with dementia, is to exclude them from norms. Many effort studies have excluded patients with dementia in part because of their generally lowered specificity rates and the fact that base rates of malingering are very low, with as few as two percent of litigants and those seeking other forms of compensation alleging vascular dementia (Mittenberg, Patton, Canyock, & Condit, 2002). As a result, the efficacy of many PVTs as they apply to dementia samples are largely unknown. Complicating matters further is the fact that if neuropsychological impairment is sufficiently severe, as in late stage dementia, patients might fail effort measures despite putting forth adequate effort (Teichner & Wagner, 2007). It is thus unclear whether many effort measures can be reliably used in this context. Little is known about which measures provide the lowest rate of false positive errors, how impairment severity and false positive rates interact, and the extent to which adjusted cut scores for dementia groups are needed (Dean et al., 2009). Several studies have examined pass/fail rates of EPVTs in dementia samples with use of measures contained within the WAIS (Wechsler, 1981; Wechsler 1997; Wechsler, 2008). Given the focus of this project, an examination of the data on the measures derived from the WAIS at various levels of impairment within dementia groups is examined.

Merton, Bossink, & Schmand (2007) examined a heterogenous group of 48 inpatients with and without “obvious” cognitive symptoms. Several standard instruments were used that included EPVTs such as the WAIS-R Digit Span subtest (Wechsler, 1981) in addition to a number of other EPVTs and SPVTs. Notably, approximately 2/3 of those taking the RDS in the “obvious” impaired group failed this measure, though variable pass rates were noted for other measures including stand-alone measures. A second experiment was run by this group looking at 20 outpatients with Alzheimer’s disease with a mean age of 73.5 and a mean score on the Mini Mental Status Exam (MMSE) performance of 22.2 compared to 14 in elderly controls with a mean age of 76.6 and a mean MMSE performance of 28.9. 70 percent. Additionally, the Alzheimer patients failed the RDS measure, whereas 64% of the elderly controls passed the RDS. Their conclusions were that the RDS may be impacted by cognitive impairment. Obvious considerations include the fact that performing adequately on a PVT with high sensitivity to detect poor effort gives one important information, although failure on the PVT may not provide much information aside from that related to false positives.

According to a study by Iverson and Tulsky (2003), suppressed Digit Span performance has been proposed as a potential marker for deliberately poor performance in a neuropsychological evaluation. The purpose of this study was to document Digit Span performance patterns in the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997) standardization sample and selected clinical groups. Base rate tables were generated for the Digit Span scaled score, longest span forward, longest span backward, and the Vocabulary–Digit Span difference score. Cut-off scores for suspecting negative response bias were proposed, and clinical case examples were used to illustrate these scores. “Based on the study results, the following guidelines are suggested for suspecting the possibility of negative response bias in an individual patient: (a) scaled score of 5, 4, or less; (b) longest span forward of 4 or less (for persons under age 55); (c) longest span backward of 2 or less; or (d) Vocabulary–Digit Span difference score of 5 or 6 (or greater)” (Iverson et al., 2003, p. 7-8). The indices discussed in this study are based on base rates in the general population and in the clinical groups that occur, for the most part, in approximately 5 percent or less of the subjects. According to Iverson and Tulsky (2003), it would be a mistake for clinicians to rely on Digit Span as their primary method for identifying biased responding. Although the specificity of the above mentioned cut-off scores is high, the sensitivity is believed to be moderate, at best. Therefore, many individuals who are exaggerating their deficits will not be identified by unusual performance patterns on this test (Iverson et al., 2003).

Ruchinskas (2019) identified 290/796 individuals diagnosed with probably early Alzheimer’s dementia, 255/796 with MCI, and 161/976 with no cognitive or neurologic abnormalities. In addition to other embedded PVTs, the subjects were administered the RDS-R (WAIS-IV; Wechsler, 2008), which includes the digits forwards, digits backwards and digits sequencing trials. ANOVAs revealed group differences for the total Digit Span performance across the three groups in addition to group differences across all three tasks (digits forwards, backwards, and sequencing). Notably, when the forward task was utilized as a covariate in MANCOVA analyses, the digits backwards and digits sequencing tasks declined by level of impairment across the groups. The authors suggest that factors related to working memory and possibly cross-task preservation (i.e. executive dysfunction) may contribute to differences across the groups, particularly with increasing level of impairment, though some preservation was noted even in the less cognitively impaired groups.

Dean et al., (2009) examined archival data from 214 dementia patients. Data was obtained from two samples consisting of a mixed dementia group and a group diagnosed with Alzheimer’s disease. Subjects had completed portions of the WAIS-R and WAIS-III to allow for calculation of the Digit Span Age-Corrected Scaled Score (ACSS), RDS, and the Vocabulary Scaled Score minus Digit Span scale score in addition to several other embedded and stand-alone PVTs. Across all patients with mean MMSE scores of approximately 20/30 correct, specificities were poor for the ACSS and RDS at 73% and 70%, respectively and 97% for the Vocabulary – Digit Span measure. Across impairment severity, the only WAIS measure that maintained adequate specificity was the Vocabulary – Digit Span measure. A similar pattern emerged relative to the specificity by type of dementia with the Vocabulary – Digit Span measure maintaining specificity above 90%. Additionally, specificity declined as performance on the MMSE declined with those obtaining a performance > 20 failing 36% of the effort measures and those with <15 correct on the MMSE failing 83% of the measures. While one may assume that those scoring less than 15 on the MMSE may be easier to classify as suffering from dementia, regardless of their effort scores, it is the subgroup with scores >20 who fail some of these measures that is likely the most relevant to clinicians.

Another factor to consider when determining if a patient’s performance is suboptimal is to look at the level of impairment. Research has shown that individuals with mild levels of impairment are able to function better on PVTs than individuals experiencing severe cognitive impairment. In a study by Kiewel et al. (2012), they examined data from 142 patients that were classified into three impairment severity groups based on their neurologic impairment, functional difficulties and neurocognitive test performance forming mild, moderate and severe groups. Measures from the WAIS included the RDS, longest digit forwards and longest digit backwards, and Vocabulary – Digit Span. Across the three severity groups, specificity remained above 90 percent for all measures within the mild group. Conversely, in the moderate group, only the longest digit forwards and Vocabulary– Digit Span remained above 90 percent. None of the measures had specificities above 61 percent in the severe group (Vocabulary – Digit Span unable to be calculated as Vocabulary was typically not administered to the severe dementia group). Only the Vocabulary – Digit Span remained above 90 percent across the entire sample. Conclusions drawn from this study include recognizing the potential value of the Vocabulary-Digit Span task, though its efficacy in those with severe dementia is largely unknown. Additionally, some indices such as the RDS may be clinically useful in the mild dementia groups, though false positives increase with dementia severity.

Limitations of Current Research with Dementia Groups

Several limitations or gaps in current research should be noted. First, due to the retrospective nature of most studies, the incremental validity of the PVTs have not been compared relative to other established measures of response bias. Secondly, the data for many of the studies were normed on populations in other countries, where many cultural variables may differ greatly from western culture. Third, in the past few years, many examinees have been educated about the procedures measuring validity, therefore ability to detect response bias has declined. Therefore, the results that were obtained for that sample may be overstating the sensitivity of the new measure. Fourth, it should also be noted that some of the samples were comprised of individuals who spoke English as their second language and differed significantly in terms of age and education (Poreh et al., 2017, p. 545).

Another limitation of current research given the heterogeneous presentation of dementia, many studies include mixed etiologies in their samples and differing diagnostic criteria. Additionally, many of these studies have been conducted utilizing a small sample size. Studies that have examined the specificity of many symptom validity tests with dementia samples have utilized cut-scores that were originally developed and normed using other patient populations. Additionally, many samples were heterogeneous with regard to neurological condition, which introduces the possibility of differing predictive validity values for the different neurological groups (Miele et al., 2012, p. 20).

Finally, one other significant limitation of current research is in the use of the published clinical norms to assess the EPVTs. Clearly, to provide more accurate results, the cutoff values should be applied to each individual case and then aggregated. Unfortunately, due to lack of access to clinical norms, the method utilized in this study provides an initial estimate of the validity of this embedded index. Additional studies using archival or new samples are needed to replicate their findings (Poreh, et al., 2017).

This literature review highlights the need for continued research into performance validity measures for individuals with neurocognitive impairment, due to the well-established fact that even when motivation is adequate, a large proportion of patients with dementia fail effort measures and are at risk for being misclassified as malingering (Bortnik et al., 2013). The purpose of this research is to examine EPVTs and identify the measures that produce the least number of false positives, with the hope that continued research can focus on further development of valid performance validity measures in dementia populations. Despite the proliferation of research on performance validity over the past decade, particularly in populations with a history of a traumatic brain injury, relatively few studies have examined the performance of individuals with dementia on commonly used embedded performance validity tests (Camara, Nathan, & Puente, 2000; Dean, Victor, Boone, Philpott, & Hess, 2009).

Finally, this research will expand the level of understanding in neuropsychological evaluation, consideration of test selection, and on studies looking at EPVTs and the criteria used with patients diagnosed with neurocognitive impairment at varied levels of impairment, by re-examining predictive factors, other than performance, that impact validity determinations. This is retrospective review of data collected from an outpatient neuropsychology clinic to evaluate EPVTs in patients with neurocognitive impairment, using the Wechsler Adult Intelligence Scales- Fourth-Edition (WAIS-IV; Wechsler, 2008).

Hypothesis #1: The WAIS-IV Digit Span subtest do not consistently detect invalid responding among different levels of cognitive impairment.

Chapter 2:

METHOD

Participants

A retrospective analysis of deidentified data collected through the Oklahoma University Health Science Center Neuropsychology's Clinical Data Base (OUHSC NCDB) was used to answer the research questions. The subsection of data used was approved by the Institutional Review Board (IRB) and the data was de-identified prior to being separated from database. No identified data was used in this analysis. Individuals included 446 individuals that previously completed neuropsychological evaluation using select embedded performance validity measures. Of the 446, all were included in the analysis.

Sample Characteristics

The sample consisted of 446 individuals of varied ethnic backgrounds, with those identifying as White or non-Hispanic comprising the largest proportion. 94.4% of individuals identified themselves as Non-Hispanic White or EuroAmerican (n = 421), 0.7% as Latino or Hispanic American (n = 3), 3.4% as African-American (n = 15), 0.9% as Native American (n = 4), and 0.4% as other (n = 2). Individuals ranged between 60 and 88 years of age, with 67 years being the median age at 4.3% (n = 19) and 68.37 being the mean age (SD = 7.03). Individuals years of education varied between 9 to 21 years, with 14 being the median years of education at 10.5% (n = 47) and 14.67 being the mean years of education (SD = 2.73). Females made up 52.2% of individuals (n = 233), males 46.9% (n = 209), and .9% of individuals did not identify their sex (n = 4). In regard to the DSM-5 Neurocognitive Diagnosis, 238 individuals were not diagnosed with a neurodegenerative disorder 54.3%, 149 individuals were diagnosed with Probable Mild Neurocognitive Disorder 33.4%, and 51 were diagnosed with Probable Major Neurocognitive Disorder 11.4%.

There were six performance validity tests (PVT) used in the analysis, all PVTs were based off of the Wechsler's Adult Intelligence Scales- Fourth Edition. DSM-5 Diagnosis were based on three diagnostic categories: No Diagnosis, Possible Mild Neurocognitive Disorder, and Possible Major Neurocognitive Disorder. De-identified data on male and female adults, ages ranging 60-88, and all racial and ethnic backgrounds were included. Individuals were referred to the OUHSC through a variety of referral sources including physicians, family members, or by self-referral. Approval from OUHSC's institutional review board was obtained for retrospective data analysis of a subset of patient's who had completed the WAIS-IV (Wechsler, 2008) from

2010-2020.The individuals that were ultimately diagnosed with a neurocognitive disorder, met the diagnostic criterion of the Diagnostic Statistical Manual-5 (DSM-5; APA, 2013). Determination of impairment severity was also made via a review of the original reports by the authors, based on reported/observed impairment in basic and/or instrumental activities of daily living, level of impairment across multiple cognitive domains.

Individuals included in the present study had no identifiable secondary gains or external incentives at the time of the evaluation. Secondary gains can range from consciously feigning poor performance to subconscious factors (e.g., fatigue due to stress about the testing process, forgetting to eat prior to testing, or effects of medication). It is common practice for the neuropsychologists at OUHSC to gather information regarding an individual’s ability to complete activities of daily living (ADL) and Instrumental activities of daily living (IADL), collateral interviews, and reviewing of medical history- which are all contributary in the evaluation process. This information is utilized when determining the preference of secondary gains. The performance of individuals in this study were found to be consistent with their presenting concerns and functional impairment, and the majority of individuals, had collateral informants who reported significant declines in cognition and instrumental activities of daily living. As a result, the following individuals included in the study were considered as having put forth adequate effort throughout the neuropsychological evaluation and any scores on effort indices indicative of malingering for individuals diagnosed with a possible mild or major neurocognitive disorder are considered false positives.

Measures

Individuals included in the present study underwent a comprehensive neuropsychological evaluation, including a clinical interview, and in most cases collateral information regarding cognition and functionality was also obtained as part of their standard care. The neuropsychological evaluations were conducted by staff clinical neuropsychologists and neuropsychology trainees at an outpatient neuropsychology clinic associated with OUHSC.

Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV)

WAIS-IV (Wechsler, 2008) is a reliable measure of intellectual abilities, with large representative normative data and extensive research indicating its high levels of validity and reliability to assess cognitive ability in individuals aged 16–90 (Pearson, 2009). The WAIS-IV has 15 subtests, 10 of which are ‘core’ and were completed in the current study. These 10 tests are scored on a scale from 1 to 19 with a mean of 10 and standard deviation of 3. The scaled scores are combined to create six composite score indices (range 50–150, Md = 100, SD = 15) derived from theoretical and factor analytic models (Wechsler, 2008). Indices include: (i) Verbal Comprehension Index (VCI), (ii) Perceptual Reasoning Index (PRI), (iii) Working Memory Index (WMI), (iv) Processing Speed Index (PSI), (v) General Ability Index (GAI), and (vi) Full-Scale IQ (FSIQ). Both GAI and FSIQ provide an overall estimate of an individual’s intellectual ability, but they differ in how they are derived. FSIQ is derived from the 10 cores subtests, whereas GAI is derived from only core Verbal Comprehension and Perceptual Reasoning subtests. Compared to FSIQ, GAI provides an estimate of general intellectual ability that has a reduced emphasis on working memory and processing speed. The dependent variables (DV) used to examine performance in the current study was primarily based on the subtest Digit Span with two DV including performance on the Vocabulary subtest.

Reliable Digit Span (RDS)

RDS is a commonly used embedded indicator of performance validity that is calculated from the Digit Span subtest of the age appropriate Wechsler Scale… "by summing the longest span of digits repeated without error over two trials under both forward and backward conditions" (Greiffenstein, Baker & Gola, 1994, p. 219-220). A cutoff score of ≤6 was applied to this analysis (Babikian et al., 2006; as cited in Young et al., 2012), meaning that if the individual scored a 6 or lower, they were identified as an invalid responder by the PVT.

Reliable Digit Span- Revised (RDS-R)

Reliable Digit Span- Revised was developed by Spencer, Tree, Drag, Pangilinan, & Bieliauskas, 2010 (Young et al., 2012). The RDS-R is calculated by adding the longest span on two trials of the same length on each of the forward, backward, and sequencing tasks on the Digit Span subtest. A cutoff score of ≤10 was applied to this analysis (Young et al., 2012), meaning that if the individual scored a 10 or lower, they were identified as an invalid responder by the PVT.

Vocabulary minus Digit Span Raw Score (VC-DS)

Vocabulary minus Digit Span was calculated by subtracting the raw score obtained on the vocabulary from the raw score obtained on the Digit Span subtest of the WAIS-IV. A cutoff score of ≥6 was applied to this analysis (Kiewel et al., 2012), meaning that if the individual had a difference of a 6 scaled points or greater, they were identified as an invalid responder by the PVT.

Age-corrected Scaled Scores on the Vocabulary minus Digit Span subtests (ACSS VC-DS)

Mittenberg, Theroux-Fichera, Zielinski,&Heilbronner (1995) introduced a derivative EVI, the Vocabulary minus Digit Span (VC–DS) age-corrected scaled score (ACSS). They noted that the normative VC–DS difference score was zero among credible individuals with traumatic brain injury (TBI), whereas malingerers exaggerated their deficits on the DS, but not VC subtest.

These findings were replicated by subsequent studies by independent researchers (Greve et al., 2003; Kiewel et al., 2012; Millis et al., 1998; as cited in Erdodi & Abear, 2019). Vocabulary – Digit Span difference scores were obtained using the individuals' Digit Span and Vocabulary subtest age-corrected scaled scores (Kiewel et al., 2012) A cutoff score of ≥3 was used (Erdodi & Abear 2019), meaning that if the individual obtained a difference of 3 scaled score points or greater, they were identified as an invalid responder by the PVT.

Age-corrected Digit Span (ACSS DS)

Digit Span Age-Corrected Scaled Score (ACSS) has generally shown more promise as a validity indicator in dementia samples as, unlike RDS, it adjusts for the age of the patient (Babikian et al., 2006; Dean et al., 2009; Heinly et al., 2005; Iverson & Tulsky, 2003; as cited in Kiewel et al., 2012). A cutoff score of ≤5 was applied to this analysis (Webber & Soble, 2018), meaning that if the individual scored a subtest scaled score of 5 or lower, they were identified as an invalid responder by the PVT.

Longest Digit Forward Trial 1 & Trial 2 (LDF-T1 & LDF-T2)

LDF-T1 was calculated at the longest digit span correctly recalled. LDF-T2 was calculated by the longest digit span correctly recalled on two consecutive trials, of the same length. A cutoff score of ≤4 was applied to LDFT1 and ≤3 was applied to LDF-T2 (Babikian et al., 2006; as cited in Kiewel et al., 2012), indicating that an individual that scores a 3 or lower, were identified as an invalid responder by the PVT.

Procedure

Since this study was archival, there were no experimental procedures. Data from the individuals included in this study were obtained retrospectively from archival data containing their test results and diagnostic information. Individuals were all administered the performance validity tests described in the Measures section above. Their scores and other information pertinent to this analysis were extracted from the data archive of the Oklahoma University Health Science Center Neuropsychology's Clinical Data Base (OUHSC NCDB) based on a study approved by the institutional IRB of the Oklahoma University Health Science Center. The Primary Investigator accessed the archive securely through an on-site computer terminal linked to the closed network at the study site. No consent was required by IRB, due to the nature of the deidentified data. In particular, the extracted data included diagnoses, age, sex, years of education, race, and test scores. Neuropsychological assessment data included scores from Wechsler's Adult Intelligence Scales- Fourth Edition.

Chapter 3: RESULTS

Overview of Analysis

In the sections below, I provide and describe the correlations among the various performance validity measures examined in this study. Once again, the primary measures of performance validity include Reliable Digit Span (RDS), Reliable Digit Span- Revised (RDS-R), Vocabulary minus Digit Span (VC-DS), Age-corrected Scale Scores for Vocabulary minus Digit Span (ACSS VC-DS), Age-corrected scaled score Digit Span (ACSS-DS), and Longest Digit Forward Trial 1 (LDF-T1), and Longest Digit Forward Trial 2 (LDF-T2). Next, I performed a series of binary logistic regression analyses to determine whether or not various demographic and diagnostic characteristics of the individuals in the study predicted the likelihood of not passing a given performance validity check. Individuals falling above the cut-score on a given validity check were coded as 0 = passed, whereas those failing to meet the cut-score were coded as 1 = failed. Each logistic regression included the following predictor variables: age, years of education, diagnosis severity, and sex. Sex was dummy coded as 0 =

female, 1 = male. Diagnosis severity (0 = no diagnosis, 1 = probable mild, probable major neurocognitive disorder, and 2 = possible major neurocognitive disorder) was recoded into two dummy variables". Both the 'Probable Major Neurocognitive Disorder' and 'Probable Major Neurocognitive Disorder' were coded as 1, whereas the baseline category ('No Diagnosis') was coded 0 for these two variables. Correlational analyses were conducted between RDS, RDS-R, VC-DS, ACSS VC-DS, ACSS DS, LDS- T1 and LDS T2 scores.

Correlations among the Performance indicators

Table 1. (see below) contains the correlations among the validity-check measures. The upper triangle of the correlation matrix contains the correlations among the raw scores, whereas the lower triangle provides correlations (Phi-coefficients) among the measures after dichotomizing based on the cut-scores for each. As expected, a number of the correlations among the scaled validity-check measures were large and statistically significant, suggesting that they were largely measuring the same thing. Nevertheless, several remarkably low and even negative correlations emerged between several of the validity-check measures. Notably, the weakest and/or most theoretically incongruent correlations emerged between the VC-DS RAW and the remaining measures and the ACSS VC-DS and the remaining measures. The strongest correlation involving the VC-DS and ACSS VC-DS emerged between themselves (where r = .842).

The lower triangle contains phi-coefficients (i.e., Pearson's correlations computed with dichotomous variables) among the dichotomized measures. A positive correlation indicates that a person who was classified as faking on one measure was more likely to be classified as faking on another. A negative correlation indicates that a person identified as faking on one measure was less likely to be classified as faking on the other. A correlation of zero indicates no relationship between the cut-score measures. Given these variables represented dichotomization of the raw scores, it is unsurprising that many of the correlations observed here are lower than those in the upper triangle. Nevertheless, similar patterns emerged in the data – where the VC-DS and ACSS VC-DS exhibited either weak and or theoretically inconsistent relationships with the other performance validity measures.

Table 1. Correlations

		RDS	RDS-R	VC-DS RAW	ACSS VC-DS	ACSS DS	LDF-T1	LDF-T2
RDS	Pearson Correlation Sig. (2-tailed) N	1 446	0.874*** 0.001 446	-0.098* 0.038 446	-0.481*** 0.001 446	0.845*** 0.001 446	0.697*** 0.001 446	0.862*** 0.001 446
RDS-R	Pearson Correlation Sig. (2-tailed) N	0.569*** 0.001 446	1 446	-0.116* 0.014 446	-0.521*** 0.001 446	0.909** 0.001 446	0.602*** 0.001 446	0.716*** 0.001 446
VC-DS RAW	Pearson Correlation Sig. (2-tailed) N	-0.117* 0.013 446	-0.048 0.310 446	1 446	0.842*** 0.001 446	-0.140** 0.003 446	-0.065 0.169 446	-0.073 0.123 446
ACSS VC-DS	Pearson Correlation Sig. (2-tailed) N	0.123** 0.009 445	0.290*** 0.001 445	0.348*** 0.001 445	1 446	-0.591*** 0.001 446	-0.396*** 0.001 446	-0.405*** 0.001 446
ACSS DS	Pearson Correlation Sig. (2-tailed) N	0.559*** 0.001 444	0.588*** 0.001 444	-0.07 0.138 444	0.236*** 0.001 443	1 466	0.706*** 0.001 446	0.721*** 0.001 446
LDF-T1	Pearson Correlation Sig. (2-tailed) N	0.439*** 0.001 446	0.282*** 0.001 446	-0.037 0.432 446	0.246*** 0.001 445	0.425*** 0.001 444	1 446	0.777*** 0.001 446
LDF-T2	Pearson Correlation Sig. (2-tailed) N	0.550*** 0.001 446	0.332*** 0.001 446	-0.099* 0.036 446	0.090 0.059 445	0.469*** 0.001 444	0.470*** 0.001 446	1 446

Notes: *** p≤.001, **p≤.01, *p≤.05. Correlations in the upper triangle are Pearson's correlations among the raw score performance validity measures. Correlations in the lower triangle are Phi coefficients, indicating the level of congruence between measures with respect to passing or failing the PVT cutoff scores. For the latter correlations, a person passing a PVT was coded 0 and failing was coded 1.

Table 2. (see below) contains descriptive statistics (e.g., mean, median, mode, and SD) for the WAIS-IV Digit-span related PVTs (RDS, RDS-R, VC-DS, ACSS VC-DS, ACSS DS, LDF-T1, and LDF-T2). The mean, median, mode, and SD were all calculated using SPSS.

Table 2. WAIS-IV Digit Span Related PVT's descriptive statistics

WAIS-IV Digit-span Related PVT	N	Mean	Median	Mode	SD	Minimum	Maximum
RDS	446	8.64	8.00	8.00	1.96	4.00	17.00

RDS-R	446	12.41	12.00	13.00	3.02	0.00	23.00
VC-DS Raw	446	14.41	15.00	16*	8.88	-10	36
ACSS VC-DS	446	1.45	1.00	1*	2.95	-7	10
ACSS DS	446	8.87	9.00	8.00	2.89	1.00	19.00
LDF- T1	446	6.13	6.00	6.00	1.29	3.00	9.00
LDF-T2	446	5.32	5.00	5.00	1.21	2.00	9.00

*Multiple modes exist- the smallest value is shown

Logistic regression: Predicting WAIS-IV VC-DS (raw score)

I performed a binary logistic regression on the WAIS-IV VC-DS embedded validity measure, where a cutoff score of ≥6 (Babikian et al., 2006; as cited in Young et al., 2012) was treated as an indicator of ‘failing’ the test. The outcome variable was coded 0 = passed and 1 = failed the measure. As noted earlier, the predictors in the model included age, years of education, diagnosis severity, and sex. Overall, the model appeared to fit the data. The chi-squared goodness of fit test was statistically significant, χ^2(df 5) = 48.654, p = <.000, indicating that the full model fit the data better than the null (no predictors) model. The Cox & Snell R-square and Nagelkerke R-square suggest that the predictors explain between 10.6% and 17.2% of the variance. The Hosmer & Lemeshow chi-square test is not statistically significant, χ^2(8)=9.036, p=.339, which is consistent with the assumption of good model fit. The overall base rate from the model was 81.6%.

Table 3. (see below) contains the regression coefficients, odds, ratios, and significance tests for the predictors in the model. Of the independent variables, only the Probable mild (neurocognitive) diagnosis variable and year of education variables were statistically significant when predicting the likelihood of failing the performance validity test. The positive regression slope (b = 1.807, s.e. = .403, p<.007 for Probable mild (neurocognitive) diagnosis indicates that persons diagnosed with a mild disorder was more likely to be classified as failing as compared to those individuals who fell into the ‘non-probable diagnosis’ group (the reference category). The Odds ratio for the ‘Probable mild’ variable indicated that the odds of failing the performance validity test for individuals with probable mild diagnosis was 2.966 times greater than that of individuals in the “no diagnosis’ category. The positive slope for years of education (b = .309, s.e. = .058, p<.001) indicates that individuals with greater levels of education were more likely to fail the PVT than those with less education. The odds ratio for this variable indicates that the odds of a person being identified as failing the validity check increased by a factor of 1.362 for each unit increment on this variable.

Sex (coded 0=female, 1=male) was a near-significant predictor in the model (b = -.469, s.e. = .269, p = .082. The negative regression slope indicates that males were less likely to be considered to have failed the validity check than females. The odds of a female failing the test was 1/.626 = 1.597 times that of males.

Table 3. WAIS-IV VC-DS

Predictor	B	SE(b)	P-value	OR
Age	-0.013	0.019	0.506	0.987
Sex (male=1)	-0.469	0.269	0.082	0.626
Years education	0.309	0.058	0.001	1.362
Diagnosis (probable mild)	1.087	0.403	0.007	2.966
Diagnosis (probable major)	0.305	0.392	0.438	1.356
Constant	-2.413	1.557	0.121	0.090

Logistic regression: Predicting WAIS-IV ACSS VC-DS

A binary logistic analysis on the WAIS-IV ACSS VC-DS embedded validity measure at a cutoff score of ≥ 6 (Erdodi & Abear, 2019). The chi-squared model test was statistically significant, χ^2(df 5) = 28.708, p = <.000. This indicates that the full model fits the data better than the null (no predictors) model. The Cox & Snell R-square and Nagelkerke R-square suggest that the predictors explain between 6.4% and 8.7% of the variance. The Hosmer & Lemeshow chi-square test is not statistically significant, χ^2(8)= 6.200, p=.625, which is consistent with the assumption of good model fit.

Table 4. (see below) contains the regression slopes, standard errors, p-values, and odds ratios for the predictors in this model. Of the independent variables, age and years of education emerged as significant predictors. The negative slope for age (b = -.043, s.e.= .016, p = .007) indicates that older individuals were less likely to fail this validity check than younger individuals. In fact, the odds ratio indicates that for every passing year, the odds changed by a factor of .958 (i.e., they were decreasing). The positive slope for years of education (b = .175, s.e. = .040, p<.001) indicates that individuals with greater levels of education were more likely to fail the PVT than those with less education.

Table 4. WAIS-IV ACSS VC-DS

Predictor	B	SE(b)	P-value	OR
Age	-0.043	0.016	0.007	0.958
Sex (male=1)	-0.092	0.209	0.661	0.912
Years education	0.175	0.040	0.001	1.192
Diagnosis (probable mild)	-0.504	0.335	0.132	0.604
Diagnosis (probable major)	-0.197	0.345	0.567	0.821

Constant	0.221	1.261	0.861	1.247

Logistic regression: Predicting WAIS-IV RDS

A binary logistic analysis on the WAIS-IV RDS embedded validity measure at a cutoff score of ≤6 (Babikian et al., 2006; as cited in Young et al., 2012), found the chi-squared model test was statistically significant, χ^2(df 5) = 16.914, p = <.005. Thus, indicating that the full model fits the data better than the null (no predictors) model. The Cox & Snell R-square and Nagelkerke R-square suggest that the predictors explain between 3.8% and 7.9% of the variance. The Hosmer & Lemeshow chi-square test is not statistically significant, χ^2(8)= 10.942, p = .205, which is consistent with the assumption of good model fit.

Table 5. (see below) contains the regression coefficients, odds, ratios, and significance tests for the predictors in the model. Of the independent variables, only the Probable mild (neurocognitive) diagnosis variable and years of education variables were significant when predicting the likelihood of failing the performance validity test. Individuals diagnosed as having a probable mild (neurocognitive) diagnosis were less likely to be identified as failing the test than those with no probable diagnosis (b = 1.393, s.e. = .450, p = .002). In fact, the odds of failing for a person not having a probable diagnosis was 1/.248 = 4.032 times that of a person with a probable mild disorder. The negative slope for years of education (b = -.124, s.e. = .061, p<.043) indicates that individuals with greater years of education were less likely to fail the PVT than those with fewer years of education.

Table 5. WAIS-IV RDS

Predictor	B	SE(b)	P-value	OR
Age	0.006	0.023	0.812	1.006

Sex (male=1)	-0.358	0.339	0.291	0.699
Years education	-0.124	0.061	0.043	0.883
Diagnosis (probable mild)	-1.393	0.450	0.002	0.248
Diagnosis (probable major)	-0.729	0.431	0.091	0.482
Constant	0.272	1.891	0.886	1.312

Logistic regression: Predicting WAIS-IV RDS-R

A binary logistic analysis on the WAIS-IV RDS-R embedded validity measure at a cutoff score of ≤10 (Young et al., 2012). Indicating the chi-squared model test was statistically significant, χ^2(df 5) = 64.483, p = <.000, indicating that the full model fits the data better than the null (no predictors) model. The Cox & Snell R-square and Nagelkerke R-square suggest that the predictors explain between 13.8% and 20.7% of the variance. The Hosmer & Lemeshow chi-square test is not statistically significant, χ^2(8)= 6.552, p = .586, which is consistent with the assumption of good model fit.

Table 6. (see below) contains the regression coefficients, odds, ratios, and significance tests for the predictors in the model. Of the independent variables, the Probable mild (neurocognitive) diagnosis (b = -2.654, s.e., = .368, p = .001) and Probable major (neurocognitive) diagnosis (b = -1.426, s.e. = .346, p = .001) variables were statistically significant when predicting the likelihood of failing the PVT. Given the regression slopes for both predictors was significant, the results indicate that persons identified as having either a probable mild deficit or a major deficit were less likely to be identified as failing the validity check than those with no probable diagnosis. The odds of a person not having a probable neurocognitive condition failing the PVT was 1/.070 = 14.286 times greater than that of a person identified as having a mild condition and 1/.240 = 4.167 times that of a person identified as having a major condition.

Table 6. WAIS-IV RDS-R

Predictor	B	SE(b)	P-value	OR
Age	0.004	0.018	0.840	1.004
Sex (male=1)	0.063	0.252	0.803	1.065
Years education	-0.017	0.046	0.711	0.983
Diagnosis (probable mild)	-2.654	0.368	0.001	0.070
Diagnosis (probable major)	-1.426	0.346	0.001	0.240
Constant	0.553	1.430	0.699	1.738

Logistic regression: Predicting WAIS-IV ACSS DS

A binary logistic analysis on the WAIS-IV ACSS DS embedded validity measure at a cutoff score of $\leq$5 (Webber & Soble, 2018).The chi-squared model test was statistically significant, χ^2(df 5) = 44.482, p = <.000, indicating that the full model fits the data better than the null (no predictors) model. The Cox & Snell R-square and Nagelkerke R-square suggest that the predictors explain between 9.8% and 19.0% of the variance. The Hosmer & Lemeshow chi-square test is not statistically significant, χ^2(8) = 3.368, p = .909, which is consistent with the assumption of good model fit.

Table 7. (see below) contains the regression coefficients, odds, ratios, and significance tests for the predictors in the model. Three of the independent variables were statistically significant in this model: Probable mild (neurocognitive) diagnosis (b = -2.713, s.e. = .450, p = .001), Probable major (neurocognitive) diagnosis (b = -1.191, s.e. = .386, p = .002), and age (b =

-.063, s.e. = .025, p = .003). An examination of the odds ratios revealed that the odds of failing the PVT for persons identified as not having a probable diagnosis was 1/.027 = 37.037 times greater than that for persons with a mild diagnosis and 1/.304 = 3.289 times that for persons with a probable major diagnosis. In effect, persons identified as having either a probable mild diagnosis or a probable major diagnosis were less likely to fail the PVT than those having no probable diagnosis. Moreover, older individuals were less likely to fail the PVT than younger individuals. Finally, the negative regression slope for the age variable indicates that older individuals were less likely to fail the PVT than those who were younger.

Table 7. WAIS-IV ACSS DS

Predictor	B	SE(b)	P-value	OR
Age	-0.063	0.025	0.013	0.939
Sex (male=1)	-0.207	0.328	0.528	0.813
Years education	-0.077	0.059	0.193	0.926
Diagnosis (probable mild)	-2.713	0.450	0.001	0.027
Diagnosis (probable major)	-1.191	0.386	0.002	0.304
Constant	5.021	2.002	0.012	151.514

Logistic regression: Predicting WAIS-IV LDF-T1

A binary logistic analysis on the WAIS-IV LDF-T1 embedded validity measure at a cutoff score of ≤ 4 (Kiewel et al., 2012). Which indicates that the chi-squared model test was not statistically significant, χ^2(df 5) = 10.208, p = <.070, indicating that the predictor variables are not moderators on if the individual's passes or fails the PVT. The Cox & Snell R-square and Nagelkerke R-square suggest that the predictors explain between 2.3% and 5.6% of the variance.

The Hosmer & Lemeshow chi-square test is not statistically significant, $\chi^2(8) = 5.414$, p = .713, which is consistent with the assumption of good model fit.

Table 8. (see below) contains the regression coefficients, odds, ratios, and significance tests for the predictors in the model. Of the independent variables, only the Probable major (neurocognitive) diagnosis variable was significant (b = 1.138, s.e. = .517, p=.028) when predicting the likelihood of failing the PVT. Specifically, probable major diagnosis was a positive predictor of the likelihood of failing relative to the non-probable diagnosis group.. The odds of failing the PVT for individuals with probable major diagnosis was 3.121 times greater than that of individuals in the "no diagnosis' category.

Table 8. WAIS-IV LDF-T1

Predictor	B	SE(b)	P-value	OR
Age	-0.029	0.029	0.318	0.972
Sex (male=1)	-0.208	0.377	0.582	0.812
Years education	-0.140	0.068	1.040	0.869
Diagnosis (probable mild)	0.596	0.426	0.162	1.814
Diagnosis (probable major)	1.138	0.517	0.028	3.121
Constant	1.136	2.122	0.593	3.113

Logistic regression: Predicting WAIS-IV LDF T2

A binary logistic analysis on the WAIS-IV LDF-T2 embedded validity measure at a cutoff score of ≤ 3 (Kiewel et al., 2012). The chi-squared model test was statistically significant, χ^2(df 5) = 420.848, p = <.001, indicating that the full model fits the data better than the null (no predictors) model. The Cox & Snell R-square and Nagelkerke R-square suggest that the predictors explain between 4.7% and 18.1% of the variance. The Hosmer & Lemeshow chi-square test is not statistically significant, $\chi^2(8) = 8.191$, p = .415, which is consistent with the assumption of good model fit.

Table 9. (see below) contains the regression coefficients, odds, ratios, and significance tests for the predictors in the model. Of the independent variables, the Probable mild (neurocognitive) diagnosis variable was a significant (b = -2.752, s.e. = .890, p = .002) predictor, along with years of education (b = -.238, s.e = .108, p = .028) when predicting the likelihood of failing the performance validity test. Individuals identified as having a probable mild diagnosis were less likely to fail the validity check than those with no probable diagnosis. In fact, those with no diagnosis were 1/.064 = 15.625 times more likely to fail the test than those with a mild diagnosis. With respect to education level, it appears that persons with more education were less likely to fail the validity check than those with less education.

Table 9. WAIS-IV LDF-T2

Predictor	B	SE(b)	P-value	OR
Age	-0.046	0.042	0.277	0.955
Sex (male=1)	-0.715	0.609	0.241	0.489
Years education	-0.238	0.108	0.028	0.788
Diagnosis (probable mild)	-2.752	0.890	0.002	0.064
Diagnosis (probable major)	-0.661	0.610	0.279	0.516
Constant	4.557	3.382	0.178	95.306

Chapter 4: DISCUSSION

The aim of this study was to examine the role of individual’s characteristics, mainly severity of impairment (measured by DSM-5 Diagnosis), in predicting the likelihood of an individual failing a performance validity test (PVT). This study examined embedded PVT using the WAIS-IV Digit Span-derived embedded measures of effort (RDS, RDS-R, VC-DS, ACSS VC-DS, ACSS DS, LDF-T1, & LDF-T2). In an effort to expand the level of understanding in neuropsychological evaluation, consideration of test selection, studies looking at EPVTs, and the criteria used with patients diagnosed with neurocognitive impairment at varied levels of impairment, this study examines the consistency of PVTs that were all derived from the same subtest and set of scores. The consideration of moderating factors in determining performance validity is a major component of neuropsychology evaluation. As the primary goal of this analysis was to examine if the individual’s level of impairment impacted PVT determination, an unanticipated finding was the inconsistency among PVTs based on the same scores obtained by individuals that took the WAIS-IV Digit-Span subtest.

In regard to correlations among the validity indices, the correlation results indicate that there are inconsistencies in what is being measured with these various PVT's. Although the high correlations among some of the PVT's appear to provide convergent validity evidence, the correlations between the VC-DS RAW and ACSS VC-DC and the remaining PVTs tended to be either very low and/or to exhibit relationships counter to what one would expect if all of the PVTs are expected to be measuring a person's performance (i.e., failing).

The logistic regression results provide further evidence that several of the PVT's may be measuring different things. Table 10. (see below) demonstrate the significant predictors in each model with the corresponding signs (direction of relationship). The table highlights the supposition that persons with worse symptoms are more likely to fail validity tests, one of the primary goals of the study, was not evident across the board. Indeed, the results lack consistency among the indicators as well as the predictor variables. In fact, this supposition is only supported if using the VC-DS raw and LDF-T1 validity indicators. It is also noteworthy that the 'possible mild' variable is more predictive with the VC-DS raw, whereas the 'possible major' variable is more predictive with the LDF-T1, demonstrating the lack of consistency in terms of the levels of severity across the two measures. On the other hand, it appears that persons considered as unlikely to have a diagnosis is more likely to fail the following PVT's (relative to the mild and/or major groups): WAIS-IV RDS, RDS-R, ACSS DS, LDF-T2. These findings do seem to provide evidence supporting the use of these measures, as a means of reasonably identifying more significant cases that are unlikely to reflect invalid responding.

Table 10. Significant Predictors in Each Model

WAIS-IV PVT	Predictor	Predictor	Predictor
RDS	Years of education (-)	Probable mild (-)	
RDS-R	Probable mild (-)	Probable major (-)	
VC-DS Raw	Years of education (+)	Probable Mild (+)	
ACSS VC-DS	Years of education (+)	Age (-)	
ACSS DS	Probable mild (-)	Probable major (-)	Age (-)
LDF-T1	Probable major (+)		
LDF-T2	Years of education (-)	Probable mild (-)	

Other variables that were found to be related to the likelihood of failing PVTs: years of education & age. Once again, however, these relationships were not consistent across the PVT measures - seemingly providing evidence of differential measurement of poor effort. Persons with greater years of education were more likely to fail the VC-DS raw and the ACSS VC-DS than those with fewer years of education, however, they were less likely to fail on the RDS and LDF - T2. Older individuals were more likely to fail the VC-DS and less likely to fail the ACSS DS.

The assigned cutoff scores were based on current research finding indicating the most appropriate scores to utilize in populations diagnosed with various neurocognitive disorders. Previous research has suggested using an RDS cutoff score of ≤7 in most clinical groups and ≤6 in various groups including individuals previously diagnosed with cerebrovascular events, severe memory disorders, and borderline intellectual functioning to assess performance validity (Schroeder, Twumasi- Ankrah, Baade, & Marshall, 2012; as cited in Mondelli, 2018). For the purpose of this study, a cutoff score of ≤6 was used on the RDS to better address the sensitivity and specificity issues in demented populations (Babikian et al., 2006; as cited in Young et al., 2012). The RDS-R is calculated by adding the longest span on two trials of the same length on each of the forward, backward, and sequencing tasks on the Digit Span subtest. A cutoff score of ≤10 was applied to this analysis (Young et al., 2012).

Research by Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner (1995) introduced the Vocabulary minus Digit Span (VC–DS) age-corrected scaled score (ACSS). They noted that the normative VC–DS difference score was zero among credible patients with traumatic brain injury (TBI), whereas malingerers exaggerated their deficits on the DS, but not VC subtest. These findings were replicated in subsequent studies by independent researchers (Greve et al., 2003; Kiewel et al., 2012; Millis et al., 1998; as cited in Erdodi & Abear, 2019). Interestingly, this analysis found that age was a significant predictive variable in the EVPTs that were adjusted for age (ACSS DS and ACSS VC-DS). Indicating that adjusting for age is a significant factor in determining invalid responding by a patient. On the VC-DS PVT research indicated that a cutoff score of ≥6 was most appropriate for a cognitively impaired patients (Kiewel et al., 2012). On the ACSS DS-VC, a cutoff score of ≥3 was used (Erdodi & Abear 2019). On the ACSS DS a cutoff score of ≤5 was found to be most robust to cognitive impairment while identifying noncredible performance (Webber & Soble, 2018). Finally, in a study by Kiewel et al. (2012), a cutoff score of ≤4 was applied to LDF-T1 and ≤3 was applied to LDF-T2 in cognitively impaired patients (Babikian et al., 2006; as cited in Kiewel et al., 2012).

Limitations and future research

Limitations of the current research include but are limited to, the data utilized for this study was part of a database of individuals referred for neuropsychological assessment, whom represent a convenience sample- in that all had been referred for clinical evaluations. Future research may be more applicable to the general population if the sample is taken and then the PVT administered based off the sample of a more general population. In reference to collecting data from individuals with cognitive impairment, sampling from that specific population (e.g., a memory care facility or other type of clinic), not only those referred for evaluation may reduce any confounding effects the database may have had.

The homogeneity of race in this sample population (94.4% Caucasian) is limiting when looking at the generalizability to the entire population. Race is an important variable that has inherent implications for treatment and should be as close to the general populations as possible. When utilizing an already established database, as is necessary for a retrospective study, it is more difficult to correct for the lack of variation in that population. This is an important aspect to consideration for future research.

Neuropsychologists rely on data from test scores, reports from other providers, and co-occurring diagnoses, there is still an element of judgment that plays a role in the evaluation. For example, behavioral observations can greatly vary depending on the person completing the observation and even the rapport that is established with the individual. As a result, the exact same individual could have very different behavioral reports which can influence the determination of valid vs invalid responding. Additionally, in circumstances where only a subset (even as low as 1) of the PVT's are administered/scored, it is possible that the clinician can arrive at radically disparate judgements of a person's test scores – depending on the PVT measures actually utilized during the assessment of motivated test-taking. In short, it is possible for two different clinicians to arrive at different conclusions regarding invalid responding , depending on which PVT's are utilized in making that assessment.

Retrospective studies have also been criticized for not having any type of control or control group. In a study by Liu and Unni (2014), researchers found that a limitation of the retrospective design is the lack of central blinded adjudication of clinical events by an independent expert group that applies consistent definitions. Another limitation is that it is unable to completely assess risk factors or confounders (Liu & Unni, 2014).

Evaluation of PVTs in clinical samples involves both practical and methodological challenges. Specifically, recruiting individuals with significant cognitive impairments can be difficult, and from a methodological standpoint, assuming that all individuals in a clinical sample are providing valid effort is never a certainty. The clinical equivalent of the simulation design would assist in evaluating the ability of a prospective PVT in identifying invalid effort through facilitating an analogue malingering group, is essential and long overdue (Leighton et al., 2014).

Conclusion

In conclusion, the logistic regression and correlation analyses provided evidence that current WAIS-IV Digit Span related PVTs did not identify any consistent predictor variables including severity of cognitive impairment, for individuals failing a performance validity test. Conversely, the performance validity tests appeared to be more appropriate for patients that were not diagnosed with cognitive impairment. Additionally, the variation among the PVTs may have attributed to the possibility that the various PVTs were influenced by different predictors. Interestingly, this study also found that the WAIS-IV PVTs were measuring characteristics other than the validity of an individual’s responses.

In a study by Leighton, Weinborn, & Mayberry (2014), the current neurocognitive literature was examined which placed PVT literature in the context of neurocognitive processing theory and identified potential methodological factors to account for the significant variability they identified in classification accuracy across current PVTs. They evaluated the utility of a well-known cognitive manipulation to provide a Clinical Analogue Methodology (CAM), that is, to alter the PVT performance of healthy individuals to be similar to that of a cognitively impaired group. Initial support was found, suggesting the CAM may be useful alongside other approaches (analogue malingering methodology) for the systematic evaluation of PVTs, particularly the influence of specific neurocognitive processing components on performance. These findings of the Leighton et al., 2014 study support these research findings. Specifically, that there are several important factors that are potentially relevant to how healthy and impaired groups may perform on PVTs. Although it remains important to recognize assessment of effort as being an integral component of any neuropsychological evaluation, embedded performance validity tests have not proven resilient to the neurological sequelae observed in moderate to severely impaired dementia populations.

This research agrees with Kiewel et al. (2012), future research should emphasize on identifying the following: external incentives and potential for secondary gain, discrepancies between test data and known patterns of brain functioning, behavioral observations during the test session, unusual presentation given the patient's documented history, and collateral information collected from friends and family (Slick, Sherman, & Iverson, 1999; as cited in Kiewel et al., 2012).

Most importantly, understanding exactly which PVTs are actually measuring performance with the least amount of interaction by a moderating variable how they may affect performance must be better understood to most effectively design new PVTs, as well as interpret the variable classification accuracy seen in existing PVTs.

The results of is research are import because they highlight the variability between PVTs that were basically all related to the individual's responses on the WAIS-IV Digit-Span subtest. When considering the validity of an individual's responses among different cognitive domains (e.g. memory, visuospatial, executive functioning) it could be explained that questionable responding on one a PVT measuring a specific domain may not actually be observed on a PVT that is measuring a different domain. That is not the case when utilizing the scores derived from testing the focuses on the same domain and especially scores derived from the same subtest. The PVTs should have the same findings if they are truly measuring the same thing. This is not something that has been found to be commonly discussed in the research and is an interesting finding of this study that has implications in determining validity of testing, treatment recommendations, as well as impacts for individuals that undergo multiple evaluations to assess changes in cognition over time. Additionally, such variability in measures purported to measure the same construct remains concerning, and efforts to explicate the reasons for these differences are needed (Leighton, Weinborn, & Mayberry, 2014).

Chapter 5: Reference

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Axelrod, B. N., Fichtenberg, N. L., Millis, S. R., & Wertheimer, J. C. (2006). Detecting incomplete effort with Digit Span from the Wechsler adult intelligence scale—Third edition. *The Clinical Neuropsychologist*, *20*(*3*), 513–523.

Badenes, G. D., Casas, H. L., Cejudo, B. J. C., & Aguilar, B. M. (2008). Evaluation of the capacity to drive in patients diagnosed of mild cognitive impairment and dementia. *Neurologia, 23*(9), 575–582.

Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various Digit Span scores in the detection of suspect effort. *The Clinical Neuropsychologist, 20*, 145–159.

Bezdicek, O., Stepankova, H., Moták, L., Axelrod, B., Woodard, J., Preiss, M., … Poreh, A. (2014). Czech version of Rey auditory verbal learning test: Normative data. *Aging, Neuropsychology, and Cognition*, *21*(6), 693–721.

Boone, K. B. (2013). *Clinical practice of forensic neuropsychology*. New York, NY: Guilford.

Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist*, *23*(4), 729–741.

Boone, K. B. (Ed.). (2007). Assessment of feigned cognitive impairment: A neuropsychological perspective. New York, NY: The Guilford Press.

Bortnik, K. E., Horner, M. D, & Bachman, D. L. (2013). Performance on standard indexes of effort among patients with dementia. *Applied Neuropsychology: Adult, 20,* 233-242.

Brower, M. (2016). Performance and Symptom Validity-2 Utilizing the Digit Span Subtest (WAIS-IV) as an Embedded Symptom Validity Test. *Archives of Clinical Neuropsychology, 31*(6), 573.

Bush, S., Ruff, R., Tröster, A., Barth, J., Koffler, S., Pliskin, N., . . . Silver, C. (2005). Symptom validity assessment: Practice issues and medical necessity NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology, 20*(4), 419-426.

Chafetz, M. D., Williams, M. A., Ben-Porath, Y. S., Bianchini, K. J., Boone, K. B., Kirkwood, M. W. et al. (2015). Official position of the American Academy of Clinical Neuropsychology Social Security Administration policy on validity testing: Guidance and recommendations for change. *The Clinical Neuropsychologist*, *29*(*6*), 723–740.

Curtis, K., Greve, K., & Bianchini, K. (2009). The Wechsler Adult Intelligence Scale—III and Malingering in Traumatic Brain Injury: Classification Accuracy in Known Groups. *Assessment, 16*(4), 401-414.

Dean, A.C., Victor, T.L., Boone, K.B., Philpott, L.M., & Hess, R.A. (2009). Dementia and effort test performance. *The Clinical Neuropsychologist, 23*, 133-152.

Denning, J. (2014). The Efficiency and Accuracy of The Test of Memory Malingering Trial 1, Errors on the First 10 Items of The Test of Memory Malingering, and Five Embedded Measures in Predicting Invalid Test Performance. *Archives of Clinical Neuropsychology, 29*(7), 729-730.

Donders, J. (2020). The incremental value of neuropsychological assessment: A critical review. *The Clinical Neuropsychologist, 34*(1), 56-87.

Donders, J., & Strong, C. (2013). Does Greens Word Memory Test really measure memory? *Journal of Clinical and Experimental Neuropsychology, 35*(8), 827-834.

Duff, K., Chelune, G., & Dennett, K. (2011). Predicting Estimates of Premorbid Memory Functioning: Validation in a Dementia Sample. *Archives of Clinical Neuropsychology, 26*(8), 701-705.

Duff, K., Patton, D., Schoenberg, M., Mold, J., Scott, J., & Adams, R. (2003). Age- and education-corrected independent normative data for the RBANS in a community dwelling elderly sample. *Clinical Neuropsychology, 17*(3), 351–366.

Erdodi, L. A. & Abeare, C. A. (2019). Stronger together: The Wechsler Adult Intelligence Scale-Fourth Edition as a multivariate performance validity test in patients with traumatic brain injury. *Archives of Clinical Neuropsychology, 00,* 1-17.

Erdodi, L., Abeare, C., Lichtenstein, J., Tyson, B., Kucharski, B., Zuccato, B., & Roth, R. (2017). Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) Processing Speed Scores as Measures of Noncredible Responding: The Third Generation of Embedded Performance Validity Indicators. *Psychological Assessment, 29*(2), 148-157.

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

Golden, C., Espe-Pfeifer, J., & Wachsler-Felder, P. (2002). Neuropsychological Interpretation of Objective Psychological Tests: Critical Issues in Neuropsychology. Boston, MA: Springer US.

Green, P. (2003). Green’s Medical Symptom Validity Test. Edmonton, AB: Green’s Publishing.

Green, P. (2003). Green’s Word Memory Test. Edmonton, AB: Green’s Publishing.

Green, R., Melo, B., Christensen, B., Ngo, L., Monette, G., & Bradbury, C. (2008). Measuring premorbid IQ in traumatic brain injury: An examination of the validity of the Wechsler

Test of Adult Reading (WTAR). *Journal of Clinical and Experimental Neuropsychology, 30*(2), 163-172.

Green, P., Flaro, L., & Courtney, J. (2009). Examining false positives on the Word Memory Test in adults with mild traumatic brain injury. *Brain Injury, 23*, 741–750.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6*, 218–224.

Greve, K., & Bianchini, K. (2004). Setting empirical cut-offs on psychometric indicators of negative response bias: A methodological commentary with recommendations. *Archives of Clinical Neuropsychology, 19*(4), 533-541.

Greve, K. W., Curtis, K. L., & Bianchini, K. J. (2013). Symptom validity testing: A summary of recent research. In S. Koffler, J. Morgan, I. S. Baron, & M. F. Greiffenstein (Eds.), Neuropsychology science and practice I (pp. 61–94). New York, NY: Oxford University Press.

Greve, K., Springer, S. Bianchini, K.J., Black, F.W., Heinly, M.T., Love, J.M., Swift, D.A., & Ciota, M.A. (2007). Malingering in Toxic Exposure: Classification accuracy of reliable digit span and WAIS-III Digit Span Scaled Scores. *Assessment, 14*(1), 12-21.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference Participants. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the Neuropsychological Assessment of Effort, Response Bias, and Malingering. *The Clinical Neuropsychologist, 23*(7), 1093-1129.

Heinly, M.T., Greve, K.W., Bianchini, K.J., Love, J.M., & Brennan, A. (2005). WAIS Digit Span-based indicators of malingered neurocognitive dysfunction: Classification accuracy in Traumatic Brain Injury. *Assessment, 12*(4), 429-444.

Henry, M., Merten, T., Wolf, S. A., & Harth, S. (2010). Nonverbal medical symptom validity test performance of elderly healthy adults and clinical neurology patients. *Journal of Clinical and Experimental Neuropsychology, 32*, 19–27.

Heyanka, D. J., Thaler, N. S., Linck, J. F., Pastorek, N. J., Miller, B., Romesser, J., & Sim, A. H. (2015). A Factor Analytic Approach to the Validation of the Word Memory Test and Test of Memory Malingering as Measures of Effort and Not Memory. *Archives of Clinical Neuropsychology*, *30*(5), 369-376.

Hook, J. N., Marquine, M. J., & Hoelzle, J. B. (2009). Repeatable Battery for the Assessment of Neuropsychological Status Effort Index performance in a medically ill geriatric sample. *Archives of Clinical Neuropsychology, 24* (3), 231–235.

Iverson, G. L., & Tulsky, D. S. (2003). Detecting malingering on the WAIS-III unusual digit span performance patterns in the normal population and in clinical groups. *Archives of Clinical Neuropsychology, 18,* 1-9.

Jasinski, L. J., Berry, D. R., Shandera, A. L., and Clark. J. A. (2011). Use of the Wechsler Adult Intelligence Scale Digit Span Subtest for malingering detection: A meta-analytic review. *Journal of Clinical and Experimental Neuropsychology 33*(3): 300-14.

Jones, D., Wilkinson, R., Jackson, C., & Drew, P. (2020). Variation and Interactional Non-Standardization in Neuropsychological Tests: The Case of the Addenbrooke's Cognitive Examination. *Qualitative Health Research, 30*(3), 458-70.

Juhasz, L. Z., Kemeny, K., Linka, E., Santha, J., & Bartko, G. (2003). The use of RBANS test (Repeatable Battery for the Assessment of Neuropsychological Status) in neurocognitive testing of patients suffering from schizophrenia and dementia. *Ideggyo´gya´szati Szemle, 56*(9–10), 303–308.

Kiewel, N. A., Wisdom, N. M., Bradshaw, M. R., Pastorek, N. J., & Strutt, A. M. (2012). A retrospective review of digit span-related effort indicators in probable Alzheimer's disease patients. *The Clinical Neuropsychologist, 26:*6, 965-974.

Kraemer, L. D., Soble, J. R., Phillips, J. I., Webber, T. A., Fullen, C. T., Highsmith, J. M., . . . Critchfield, E. A. (2020). Minimizing Evaluation Time While Maintaining Accuracy: Cross-Validation of the Test of Memory Malingering (TOMM) Trial 1 and First 10-Item Errors as Briefer Performance Validity Tests. *Psychological Assessment, 32*(5), 442-450.

Kotani, S., Sakaguchi, E., Warashina, S., Matsukawa, N., Ishikura, Y., Kiso, Y., et al. (2006). Dietary supplementation of arachidonic and docosahexaenoic acids improves cognitive dysfunction. Neuroscience Research, 56 (2), 159–164.

Larrabee, G. J. (2015). The multiple validities of neuropsychological assessment. *The American Psychologist*, *70*(8), 779–788.

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*, 1–7.

Larrabee, G. T. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologists, 22*, 410-425.

Leighton, A., Weinborn, M., & Mayberry, M. (2014). Bridging the gap between neurocognitive processing theory and performance validity assessment among the cognitively impaired: A review and methodological approach. *Journal of the International Neuropsychological Society, 20*, 873-886.

Lippa, S. M., Lange, A. B., Bhagwat, A., & French, L. M. (2017). Clinical utility of embedded performance validity tests on the repeatable battery for the assessment of neuropsychological status (RBANS) following mild traumatic brain injury. *Applied Neuropsychology: Adult, 24*, 73-80.

Liu., Y., & Unni, E. (2014). Methodological issues of retrospective studies assessing health outcomes of potential clopidogrel-statin interaction. *The International Journal of Pharmacy Practice, 22*(5), 360-362.

Loring, D. W., Glenn, J. L., Lee, G. P., & Meador, K. J. (2007). Victoria symptom validity test performance in a heterogenous clinical sample. *The Clinical Psychologist, 21*, 522-531.

Mcdermott, B. (2012). Psychological Testing and the Assessment of Malingering. *Psychiatric Clinics of North America, 35*(4), 855-876.

Merton, T., Bossink, L, & Schmand, B. (2007). On the limits of effort testing: Symptom validity test and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology, 29*(3), 308-318.

Miele, A., S., Gunner, J., H., Lynch, J., K., & McCaffrey, R., J. (2012). Are Embedded Validity Indices Equivalent to Free-Standing Symptom Validity Tests? *Archives of Clinical Neuropsychology, 27*(1), 10-22.

Mittenberg, W., Azrin, R., Millsaps, C., & Heilbronner, R. (1993). Identification of Malingered Head Injury on the Wechsler Memory Scale—Revised. *Psychological Assessment, 5*(1), 34-40.

Mittenberg, W., Theroux-Fichera, S., Zielinski, R., & Heilbronner, R. (1995). Identification of malingered hHead Injury on the Wechsler Adult Intelligence Scale—Revised. *Professional Psychology: Research and Practice, 26*(5), 491-498.

Mittenberg, W., Fichera, S., Zielinski, R., & Heilbronner, R. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale-Revised. *Psychological Assessment, 5*, 34-40.

Mittenberg, W., Patton, C., Canyock, E., & Condit, D. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*, 1094–1102.

Mondelli, J., Ernst, William, Kneavel, Meredith, & Lawler, Kathy. (2018). Reliable Digit Span Performance in Individuals with Parkinson's Disease or Various Forms of Dementia, ProQuest Dissertations and Theses.

Navarro, G., Schultheis, Maria, Chute, Douglas, & Hickey, Chelsea. (2019). The Effects of Repeated Exposure on the Vocational Multitasking Test, ProQuest Dissertations and Theses.

Odland, A. P., Lammy, A. B., Martin, P. K., Grote, C. L., & Mittenberg, W. (2015). Advanced administration and interpretation of multiple validity tests. *Psychological Injury and Law, 8*, 46–63.

Paulson, D., Horner, M. D., & Bachman, D. (2015). A comparison of four validity indices for the RBANS in a memory disorders clinic. *Archives of Clinical Neuropsychology, 30,* 207-216.

Pearson. (2009). Advanced Clinical Solutions for the WAIS-IV and WMS-IV- Technical Manual. San Antonio, TX: Pearson.

Poreh, A. (2005). Analysis of mean learning of normal participants on the Rey Auditory-Verbal Learning Test. *Psychological Assessment*, *17*(2), 191–199.

Poreh, A., Tolfo, S., Krivenko, A., & Teaford, M. (2017). Base-rate data and norms for the Rey Auditory Verbal Learning Embedded Performance Validity Indicator. *Applied Neuropsychology: Adult, 24*(6), 540-547.

Prigatano, G. P., & Pliskin, N. H. (2003) Clinical Neuropsychology and Cost Outcome Research. London, England: Psychology press.

Proto, D. A., Pastorek, N. J., Miller, B. I., Romesser, J. M., Sim, A. H., & Linck, J. F. (2014). The dangers of failing one or more performance validity test in individuals claiming mild traumatic brain injury-related postconcussive symptoms. *Archives of Clinical Neuropsychology, 29,* 614-624.

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five Validation Experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment, 10*(1), 10-20.

Reynolds, C. (1982). Determining statistically reliable strengths and weaknesses in the performance of single individuals on the Luria-Nebraska Neuropsychological Battery. *Journal of Consulting and Clinical Psychology, 50*(4), 525-529.

Ryan, J. J., Glass, L. A., Hinds, R. M., & Brown, C. N. (2010). Administration order effects on The Test of Memory Malingering. *Applied Neuropsychology*, *17,* 246-250.

Schoenberg, M. R., & Scott, J. G. (2011). The Little Black Book of Neuropsychology: A Syndrome-Based Approach. Boston, MA: Springer US.

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable Digit Span: A systematic review and cross-validation study. *Assessment*, *19*(1), 21-30.

Sharma, V., Krishna, M., Lepping, P., Palanisamy, V., Kallumpuram, S., Mottram, P., . . . Copeland, J. (2010). Validation and feasibility of the Global Mental Health Assessment

Tool—Primary Care Version (GMHAT/PC) in older adults. *Age and Ageing, 39*(4), 496-499.

Silverberg, N. D., Wertheimer, J. C., & Fichtenberg, N. L. (2007). An effort index for the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). *The Clinical Neuropsychologist, 21*, 841–854.

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13,* 545-561.

Slick, Tan, J.E., Strauss, E., Mateer, C.A., Harnadek, M., & Sherman, E.M.S. (2003). Victoria Symptom Validity Test scores of patients with profound memory impairment: Nonlitigant case studies. *The Clinical Neuropsychologist, 17,* 390-394.

Slick, D. J., Tan, J. E., Strauss, E. H., & Hultsh, D. F. (2004). Detecting malingering: A survey of experts' practices. *Achieves of Clinical Neuropsychologist, 19,* 465-473.

Slick, D. J., & Sherman, E. M. S. (2013). Differential Diagnosis of Malingering. In D. A. Carone & S. S. Bush (Eds.), Mild traumatic brain injury: Symptom validity assessment and malingering (pp. 57-72). New York, NY: Springer.

Spinks, R., McKirgan, L. W., Arndt, S., Caspers, K., Yucuis, R., & Pfalzgraf, J. (2008). IQ estimate smackdown: IQ proxy measures to the WAIS-III. *Journal of International Neuropsychological Society, 15,* 590-596.

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

Teichner, G., & Wagner, M. (2004). The Test of Memory Malingering (TOMM): Normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Archives of Clinical Neuropsychology, 19*(3), 455-464.

Tombaugh, T.N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment, 9,* 260-268.

Tse, V. W., Crabtree, J., Islam, S., & Stott, J. (2019) Comparing intellectual and memory abilities of older autistic adults with typically developing older adults using WAIS-IV and WMS-IV. *Journal of Autism and Developmental Disorders*, *49*(10), 4123-4133.

Webber, T. A., Bailey, A., Alverson, K., Critchfield, C., Bain, W., Messerly, A., . . . Soble, J. (2018). Further Validation of the Test of Memory Malingering (TOMM) Trial 1 Performance Validity Index: Examination of False Positives and Convergent Validity. *Psychological Injury and Law*, *11*(4), 325-335.

Webber, T. A., Critchfield, E. A., & Soble, J. R. (2018). Convergent, discriminant, and concurrent validity of non-memory-based performance validity tests. *Assessment*, 1-17.

Webber, T.A., & Soble, J.R. (2018). Utility of various WAIS-IV Digit Span indices for identifying noncredible performance validity among cognitively impaired and unimpaired examinees. The Clinical Neuropsychologist, 32(4), 657-670.

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

Wechsler, D. (1997). Wechsler Adult Intelligence Scale (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008). Wechsler Adult Intelligence Scale-Fourth edition. San Antonio, TX: Pearson Assessment.

Young, J. C., Sawyer, R. J., Roper, B. L., & Baughman, B. C. (2012). Expansion and re-examination of digit span effort indices on the WAIS-IV. *The Clinical Neuropsychologist, 26*(1), 147-159.

Zillmer, E. A., Spiers, M. V., & Culberson, W. C. (2008). Principles of Neuropsychology, Second Edition. Belmont, CA: Thompson Wadsworth.