ASSESSMENT OF FEIGNED NEUROCOGNITIVE IMPAIRMENT IN RETIRED ATHLETES IN A MONETARILY INCENTIVIZED FORENSIC SETTING

Jesse M. Smotherman, M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2020

APPROVED:

Jennifer L. Callahan Committee Chair
Charles A. Guarnaccia, Committee Member
Craig S. Neumann, Committee Member
Nick M. Wisdom, Committee Member
Vicki Campbell, Chair of the Department of Psychology
Tamara L. Brown, Executive Dean of the College of Liberal Arts and Social Sciences
Victor Prybutok, Dean of the Toulouse Graduate School

Smotherman, Jesse M. *Assessment of Feigned Neurocognitive Impairment in Retired Athletes in a Monetarily Incentivized Forensic Setting*. Doctor of Philosophy (Clinical Psychology), August 2020, 64 pp., 8 tables, references, 97 titles.

Compromised validity of test data due to exaggeration or fabrication of cognitive deficits inhibits the capacity to establish appropriate conclusions and recommendations in neuropsychological examinations. Detection of feigned neurocognitive impairment presents a formidable challenge, particularly for evaluations involving possibilities of significant secondary gain. Among specific populations examined in this domain, litigating mild traumatic brain injury (mTBI) samples are among the most researched. One subpopulation with potential to contribute significantly to this body of literature is that of retired athletes undergoing fixed-battery neuropsychological evaluations within an assessment program. Given the considerable prevalence of concussions sustained by athletes in this sport and the substantial monetary incentives within this program, a unique opportunity exists to establish rates of feigning within this population to be compared to similar forensic mTBI samples. Further, a fixed battery with multiple validity tests (VT) offers a chance to evaluate the classification accuracy of an aggregated VT failure paradigm, as uncertainty abounds regarding the optimal approach to the recommended use of multiple VTs for effort assessment. The current study seeks to examine rates of feigned neurocognitive impairment in this population, demonstrate prediction accuracy equivalence between models based on aggregated VT failures and logistic regression, and compare classification performance of various individual VTs.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

CHAPTER 1

DETAILED REVIEW OF LITERATURE

Introduction

A marked surge in interest regarding the development of reliable, valid techniques for detecting feigned neurocognitive deficits has been observed in recent years (Heilbronner et al., 2009). This rise in demand for more precise appraisals of the presence of compromised validity has occurred in tandem with the increased integration of neuropsychologists into a wide array of forensic evaluations (Vitacco et al., 2008). Findings from neuropsychological assessments are highly influential in the decision-making processes for disability determinations, litigation proceedings, and competency cases, among others (Heilbronner et al., 2009). Common among evaluations within these categories is their shared possibility of secondary gain (e.g., financial rewards). Furthermore, some research, involving individuals with traumatic brain injury (TBI) participating in neuropsychological testing associated with varying degrees of potential monetary compensation, has demonstrated a positive correlation between financial incentive level and feigning likelihood (Bianchini, Curtis, & Greve, 2006). Such findings have underscored the need for a consensus definition of malingering, as well as more accurate assessments of symptom validity and effort measurement (Sharland & Gfeller, 2007).

These observations in the literature intimate a broader problem within neuropsychological assessment; that is, accurately pinpointing non-neurologic factors capable of skewing test performance. As purely psychological contributions (e.g., preexisting mood disorders) to the clinical picture can negatively impact an individual's assessment performance, it is important, albeit challenging, to differentiate these features from strictly neurologic determinants causing poor performance (Lezak et al., 2012). This significant challenge is further

amplified by the insidious onsets common to many neurodegenerative diseases (Kanazawa, 2001), wherein observable symptoms in early stages (e.g., headache, inattention, memory problems) overlap with many psychiatric conditions, often resulting in misdiagnoses (Bradford et al., 2009). Moreover, when evidence for brain injury is limited to neuropsychological assessment data alone, with no "concrete" collateral support from independent neurological examinations, the fundamental goal of accurate interpretability becomes all the more elusive. Better conceptual clarity of, and testing for, feigned neurocognitive impairment is highly needed.

Feigned Cognitive Impairment in Neuropsychological Assessment

Given the pervasive ambiguity apparent in many aspects of the current framework of neuropsychological assessment, the overarching pursuit of drawing accurate conclusions and making proper treatment recommendations is frequently hindered by concerns of tenuous validity (Bigler, 2012; Heilbronner et al., 2009). The added possibilities of deliberate attempts to distort history, exaggerate symptoms, or falsify responses, therefore, sharply intensify the existing challenges inherent in this pursuit. Before the 1980s, prior to the accelerated involvement of neuropsychologists in forensic evaluations, where feigned impairment is often highly relevant, research on malingering was limited (Slick & Sherman, 2012). As more concerted efforts to derive a functional definition of malingering were exerted in the latter part of the 20th century, some progress was made in developing useful techniques for detecting feigned cognitive deficits (Nies & Sweet, 1994); however, the difficulty in delineating the nuanced aspects of malingering, in conjunction with the consequential gravity of misdiagnoses (particularly regarding "high stakes" false-positives), left the field without a consensus conceptualization or standard of practice on this subject.

Due to broad disagreement among researchers about the construct of malingering and the

appropriate clinical approach for assessing effort, the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition (DSM-IV; American Psychological Association, 1994) failed to provide official diagnostic criteria for malingering (Slick, Sherman, & Iverson, 1999). In lieu of classifying malingering as a formal diagnosis, the DSM-IV proposed a malingering V-code to be assigned when individuals engage in *conscious* feigning for *external* goals. Additionally, a framework was provided for guiding differential diagnoses between several disorders in which fabricated symptoms are germane (e.g., conversion disorder, factitious disorder, somatic symptom disorder). This framework, which persists in the latest version of the DSM along with the aforementioned V-code (DSM-5; American Psychological Association, 2013), employs two dichotomous variables in distinguishing between relevant disorders (i.e., volitional [yes or no] and goals [external or intrapsychic]). Under this system, it is posited that a clinician navigates diagnostic considerations by determining whether an individual's reportedly false symptoms are feigned intentionally or subconsciously and if incentives are external (e.g., financial reward, avoiding formal duties, drug-seeking) or psychological (e.g., sick role, manage stress).

Subsequent to the introduction of this DSM model of malingering, substantial criticism emerged, largely targeting its heavy reliance on judgment calls required for subjective matters (e.g., the extreme difficulty of defining intent), as well as the minimal assistance offered for discerning whether feigned deficits are exaggerated or fabricated (Berry & Nelson, 2010). Furthermore, the dichotomization of malingering as a classification has garnered disapproval, as some research has generated support for a more dimensional structure of malingering, based on feigned psychopathology findings (Walters et al., 2008). Such results cast doubt on the suitability of a quantitative approach to diagnosing malingering.

Further complicating the process of accurately assessing the presence or degree of

feigned neurocognitive impairment is the wide variability among available base rate statistics available (Mittenberg, Patton, Canyock, & Condit, 2002). As accurate malingering identification depends on consideration of appropriate base rates (Ardolf, Denney, & Houston, 2007), erroneous or excluded base rate data can lead to misclassification (Rosenfeld, Sans, & Van Gorp, 2000). Additionally, the confounding effects of the unknown number of undetected malingerers necessarily being omitted from base rate calculations contribute to underestimation of prevalence approximations of cognitive feigning (Rosenfeld, Sans, & Van Gorp, 2000). Despite the significant variability in malingering incidence rates reported across settings and patient populations, a general trend of feigned neurocognitive impairment occurring more frequently in more incentivized evaluations is observed (Ardolf et al., 2007; Bianchini et al., 2006; Larabee, 2003).

In response to the growing need for an operational definition of malingering, as well as for diagnostic parameters, Slick, Sherman, and Iverson (1999) proposed criteria for malingered neurocognitive dysfunction (MND), with some conceptual updates later provided by Slick, Tan, Sherman, and Strauss (2010). Essential elements required for this diagnosis include evidence of the conscious expenditure of efforts to exaggerate or fabricate neurocognitive deficits (immediate goal), which serve to obtain material gain or avoid formal duty (long-term goal) (Slick & Sherman, 2012). Importantly, the Slick et al. diagnostic framework abandons the strict dichotomous approach to malingering, instead offering three graded levels of MND certainty: definite, probable, and possible. The criterion of "the presence of a substantial external incentive" is shared among all three levels (Slick, Sherman, & Iverson, 1999). For definite or probable MND, a lack of alternative explanations (e.g., psychiatric factors) for behavior or test data is also required. In addition, a diagnosis of definite MND also requires the criterion of

neuropsychological testing evidence be met by definite negative response bias, defined as either below chance ($p < .05$) performance on forced-choice measure(s), high posterior probability (> .95) on well-validated psychometric feigning indices, or discrepancies between test data and known patterns of brain functioning, observed behavior, collateral report, or history (Slick & Sherman, 2012). In contrast, a probable MND diagnosis requires at least three types of discrepancies between test data and other obtained information (e.g., collateral report, observed behavior, etc.), excluding definite negative response bias. Finally, a possible MND diagnosis is reached when some evidence of feigned cognitive dysfunction is present in the absence of alternative explanations or when definite or probable MND criteria are met, but alternative explanation possibilities cannot be ruled out (Slick & Sherman, 2012).

Much criticism of the Slick et al. (1999) framework has surfaced since its introduction and subsequent widespread clinical application. Most notably, Rogers, Bender, and Johnson (2011) spotlight the marked discrepancy between relatively lenient MND rule-in criteria, which leans heavily on inferred motivation for external reward, and relatively rigorous MND rule-out criteria. The authors argue against such discrepancy models, highlighting the vulnerability to error resultant from failure to adequately weigh a myriad of possible factors capable of contributing significantly to below-expectation test results. Yet, even with considerable inadequacies outlined in such criticisms within this body literature, the Slick et al. framework remains the most commonly recommended and utilized guideline for assessing malingering in neuropsychological practice (Heilbronner et al., 2009).

Neuropsychological Feigning Detection Methods

Beyond the problematic conceptual aspects of accurately defining the essential components of feigned neurocognitive impairment, ample consideration of the psychometric

validity of feigning detection methods is warranted. Generally, tests designed to detect feigned neurocognitive deficits are classified as either embedded or stand-alone validity measures. Embedded validity measures (EVMs) are developed for traditional neuropsychological assessment measures, whereas stand-alone validity measures (SVMs) are distinctly separate from other existing tests of neuropsychological functioning (Lezak et al., 2012). The development of such symptom validity measures has aided neuropsychologists, in that the challenge of determining the validity of neuropsychological test findings no longer falls solely on their clinical judgment (Bigler, 2014). With the injection of some objectivity into an area fraught with subjective pitfalls, symptom validity measures designed to assess performance or response validity have assisted, at least partially, in better defining and evaluating the construct of malingering.

In comparing the clinical value of EVMs and SVMs across many domains of consideration relevant to overarching and specific goals of neuropsychological assessment, Bigler (2014) sheds light on several distinct advantages of EVMs over SVMs. One obvious benefit is the removal of the need for extra, response-bias specific testing tacked onto assessment batteries. In addition to saving time and money, the reduction in total testing time also mitigates the burden of the evaluation process on the examinee, which can help lessen the influence of testing fatigue onto performance validity (Lezak, 2012). Likewise, embedding validity measures into existing tests offers the advantage of sharing construct validity with its parent test, unlike stand-alone measures, which necessarily demand that results be generalized to intended constructs validated separately by other psychometric tests (Bigler, 2014).

Conversely, unlike EVMs, SVMs are not limited by the potentially increased risk that results might be impacted by cognitive ability, as opposed to reflecting strictly effort or response

bias, EVMs are contained within tests originally designed to quantify more multifaceted properties of neurocognitive functioning (Lezak, 2012). One other remarkable advantage of SVMs is evident in specific, stand-alone forced-choice tasks, from which deliberate under-performance can be identified with statistical confidence, derived from well-supported research on significantly below chance performance base rates in certain populations (Greve, Binder, & Bianchini, 2009; Frederick & Speed, 2007). Although the sizeable amount of empirical research on SVM forced-choice measures demonstrates their utility in performance validity assessment, the common, simplistic design and widespread usage of forced-choice memory tasks put them at higher risk for compromised test security and coaching concerns (Youngjohn, Lees-Haley, & Binder, 1999).

Coaching within the context of neuropsychological testing in forensic evaluations refers to deliberate briefing of an examinee, usually by their attorneys, on the likely presence of symptom validity measures in an upcoming neuropsychological testing battery, or outright instructions on specific responses to give, both of which threaten to invalidate findings (Brennan et al., 2009). Moreover, coaching can also include advice on presenting with prominent symptoms of certain neurological disorders; however, research on specific symptom coaching suggests that this approach to test subversion is not as effective as specific test coaching (Brennan et al., 2009). Some research has suggested that rates of coaching are particularly high in personal injury evaluations involving brain injury (Essig et al., 2001). Together with increasing threats to test security stemming from dissemination of specific test information over the internet (Kaufmann, 2009), coaching in forensic neuropsychological evaluations exposes an important vulnerability to data validity preservation. As the validity of some of the most common SVM forced-choice measures might already be significantly compromised by these

threats (Ruiz et al., 2002), the need for newer, more accurate feigning detection techniques and enhanced test security methods is clear.

In refining and improving upon the design process of feigning detection measures, demonstration of appropriate sensitivity and specificity is critical to test validation. For tests with continuous independent predictor variables, sensitivity signifies the percentage of true positives identified at a given cut score. That is, in the context of malingering detection, how many malingerers were correctly classified. Similarly, a given test's specificity also indicates a percentage of correct classification at the same chosen cut score; however, specificity is concerned with true negatives (i.e., how many non-malingerers were correctly classified). Importantly, sensitivity and specificity data should be combined with respective base rates within a given patient population when gauging a measure's predictive power (Lezak, 2012).

Also key to note in addressing the validity of dichotomous models of malingering detection is a problem with cut-score-reliant measures where classification errors (e.g., false positives) more likely occur near the chosen cut score. Given the high stakes nature of many forensic evaluations in which such dichotomous models are commonly employed, the prospect of an increased likelihood of committing type I or type II errors when results are close to the cut score is particularly concerning. Rogers (1997) addressed this concern by positing the inclusion of an indeterminate category, in which scores falling within its threshold are restricted from contributing to final decision-making. While this suggestion is appealing in its error reduction quality, further research is needed to ascertain the extent to which a third category impacts the accuracy of specificity and sensitivity measurements (Rogers, 1997).

In a detailed analysis of malingering detection strategies, Rogers, Harrell, and Liff (1993) attempted to improve on these and other limitations of common methods of feigned

neurocognitive impairment detection by offering a more comprehensive conceptual framework. Specifically, the authors demarcated the focus of feigning detection into unlikely presentations (i.e., atypical patterns of responding compared to individuals with true neurocognitive impairment) and excessive impairments (i.e., significantly lower-than-expected performance when accounting for genuine neurocognitive deficits). Perhaps the most valuable contribution from the research efforts of these authors, as well as Rogers and Correa (2008), lies in the multifaceted detection strategies posited within these two symptom presentations. Specifically, by outlining a dimensional approach (e.g., emphasizing importance of magnitude of error, performance curve analysis, floor effects, etc.) for determining the extent to which unlikely presentations and excessive impairments suggest deliberate distortion of neuropsychological status, Rogers and Correa shift the focus of assessment toward degree of impairment and away from category. In assigning more weight to the measurement of nuanced aspects of feigned neurocognitive impairment, such a dimensional approach better acknowledges the complex nature of malingering, and, in turn, likely reflects a more accurate representation of the underlying forces driving this phenomenon.

In the same vein, other special considerations in diagnosing malingering echo the extraordinary challenge of defining and accurately assessing whether an individual is feigning impairment by his/her own volition for the explicit goal of procuring a substantial external reward. For example, the issue of whether warning examinees before testing of the existence of effort tests yields more sophisticated malingering (Suhr & Gunstad, 2000) or reductions in malingering (Schenk & Sullivan, 2010) remains highly debated. Likewise, the fact that neurological dysfunction and capacity to feign impairment are not mutually exclusive further muddles the clinical picture (Lezak, 2012). Similarly, preexisting psychiatric disorders,

individual personality differences, emerging “neurolaw” specialty attorneys, and patients with true impairment who seek to minimize deficits (Pankratz & Erickson, 1988) are all examples of subjects that require more concentrated research efforts due to their common capacity for producing confounding results, inaccurate interpretations, and contraindicated treatment recommendations. Given the overwhelming undertaking of elucidating malingering science as a whole, more focused investigations of specific patient populations (e.g., TBI) might prove fruitful in revealing broader truths of feigned neurocognitive impairment overall.

Feigned Neurocognitive Impairment Detection in TBI Population

Individuals who have suffered traumatic brain injury (TBI) make up the most researched comparison group within the literature on symptom validity testing in neuropsychological assessment (Bigler, 2014). In light of the fact a large percentage of incidents resulting in TBI involve injuries sustained from motor vehicle accidents, falls, and assaults (Faul & Coronado, 2015), it is logical that so many litigants and claimants referred for forensic neuropsychological evaluation report TBI-related symptomology. Similarly, as TBI has been identified as one of the leading causes of disability in the U.S., with approximately three million individuals experiencing TBI-related disability (Zaloshnja et al., 2005), it is also not surprising that neuropsychological injuries have become increasingly more common as matters of relevance within the legal system (Sweet et al., 2011). In the U.S., total estimated costs (i.e., direct [medical] and indirect [productivity loss]) associated with TBI were $76.5 billion in 2000 (Sweet et al., 2013), a significant sum that underscores a major driving force behind the striking rise in attention directed toward TBI-related research, clinical practice, and legal issues.

Interest has spiked within the body of literature pertaining to TBI in neuropsychology; specifically, research on detection strategies for TBI-related feigned neurocognitive impairment

has escalated, as rates of response bias reliably increase in forensic versus other clinical settings (Carone & Bush, 2013). Consequently, the inclusion of SVMs and EVMs into assessment batteries was recommended by the American Academy for Clinical Neuropsychology (AACN) for all neuropsychological evaluations involving the possibility for secondary gain (Heilbronner et al., 2009; Sweet et al., 2013). Notably, among the varying gradations of TBI severity, mild TBI (mTBI) stands out as particularly problematic in terms of response bias, as approximately 40% of individuals involved in compensation-seeking matters or litigation fail effort tests (Larrabee, 2007). Moreover, multiple studies have demonstrated that individuals with mTBI involved in litigation perform worse, on average, than individuals with moderate or severe TBI litigants on cognitive performance tasks (Carone, 2008; Wood & Rutterford, 2006). Given these repeated findings, the general recommendation that multiple performance validity tests be administered for cases involving secondary gain potential (as this combination significantly increases malingering classification accuracy; Greve, Binder, & Bianchini, 2009), is expressly warranted for forensic cases involving mTBI.

Despite the relatively high number of studies examining feigning detection measures and response bias in TBI samples, meaningful conclusions drawn from such studies are not always possible. Specifically, the high psychiatric comorbidity rates characteristic of patients with TBI often result in the exclusion of individuals with TBI comorbidities in feigning measure studies, which limits generalizability of results (Fann et al., 1995). Coupled with the aforementioned elevations in feigned impairment rates observed in mTBI litigants, research attempts to build upon the extant literature are often confounded and drawing broader conclusions from narrowed sample data is not always possible.

Further obscuring diagnostic clarity in this domain is the lack of a consensus definition of

which properties specifically constitute mTBI, concussion, and post-concussive syndrome (PCS) (Sweet et al., 2013). Some diagnostic criteria for mTBI depend on posttraumatic amnesia (PTA), others rely on duration of loss of consciousness (LOC), while the Glasgow Coma Scale (GCS), among the most commonly utilized instruments for assessing TBI severity, was constructed as a measure of responsiveness (Lezak, 2012). Regarding PCS, hallmark aspects involve the persistence (i.e., more than three months after injury) of mTBI-like symptoms (e.g., headache, fatigue, dizziness). Sweet et al. (2013) outlined the complicated problem of differentiating between these similar, ill-defined closed head injuries in forensic scenarios in which retrospective diagnoses are suggested far after the date of the reported injury. The chief problem in this scenario lies in the well-documented inaccuracies of self-report data (e.g., Iverson et al., 2010), which echoes the larger challenge in this area of research, in that immediate assessments of mTBI can be difficult to achieve.

Amidst the ambiguity apparent in much of the research centered on assessing and defining mTBI, the challenge of selecting the best available measures for accurately detecting feigned neurocognitive impairment is amplified, especially within a “high stakes” forensic context. To assist in this sizeable task, Berry and Schipper (2007) offer helpful quality control standards for determining which detection measures have been psychometrically validated for use in a given patient population. These standards mandate that high quality feigning detectors: have support from peer-reviewed studies on sensitivity and specificity data in a given population, provide differential predictive power values for various base rates as well as clinical comparison group data, demonstrate validity in both simulation and known-group methodologies, and have empirical support from independent researchers (Berry & Schipper, 2007). With the guidance of these standards, some EVMs and SVMs stand out as more robust and suitable for assessing

validity of neurocognitive dysfunction complaints in a TBI sample than others.

The stand-alone measures of symptom validity identified in the literature as most suitable for reliably detecting feigned impairment in TBI cases have met the quality control standards put forth by Berry and Schipper (2007), and have been extensively researched within TBI-specific samples. The basic premise upon which psychometric instruments in this domain are designed is that individuals with severe neurologic dysfunction, due to varying disorders known to cause cognitive impairment, have reliably demonstrated no difficulty in passing these SVMs with ease, as evidenced by multiple empirical studies (Breting & Sweet, 2013). Therefore, individuals with disorders classified as less severe by definition, like mTBI, should not exhibit any difficulty in passing such SVMs.

One such measure is the Word Memory Test (WMT; Green, 2005), a computerized, forced-choice, word-based SVM. The WMT has the benefit of a convincing sum of empirical evidence that demonstrates its value in forensic neuropsychological evaluations of mTBI patients, specifically. Notably, studies have shown that mTBI litigants reliably perform worse on the WMT than individuals with higher TBI severity who are not involved in litigation (Green, Iverson, & Allen, 1999). Also, WMT failure in individuals with mTBI predicted worse performance in neuropsychological assessment batteries overall compared to individuals with more severe TBI who did not fail the WMT (Green et al., 2001).

Similarly, the Test of Memory Malingering (TOMM; Tombaugh, 1996) is also a widely used, forced-choice SVM. While the TOMM is visually-based, it still relies on recognition memory (for simple pictures) and gives examinees immediate feedback on accuracy of their responses. Also like the WMT, much research exhibiting its usefulness in forensic environments has focused on mTBI samples. For example, Haber and Fichtenberg (2006) sought to replicate

the TOMM's original validation study (using 45/50 on trial 2 as a cut off score), and found that 93% of non-malingering mTBI examinees in monetarily incentivized contexts were correctly classified (good specificity), while 64% of malingering mTBI examinees were detected (moderate sensitivity). Comparable results were found in another study showing 90% specificity and 60% sensitivity rates for mTBI groups (Greve, Bianchini, & Doane, 2006).

For embedded validity indicators, forced-choice designs are also common; however, some EVMs are formatted to take advantage of floor effects (i.e., an excessive impairment detection strategy in which cognitive ability for task success above a given "floor" is preserved in populations with true neuropsychological impairment), as well as atypical response patterns (Breting & Sweet, 2013). One EVM included in all versions of the Wechsler Adult Intelligence Scale (WAIS) is Reliable Digit Span (RDS), which utilizes the Digit Span (DS) subtest of the WAIS by adding the highest number of digits correctly repeated on both trials of a given number length for DS forward and backward (Greiffenstein, Baker, & Gola, 1994). RDS is among the most researched techniques for performance validity in various clinical populations, with findings supporting its use, albeit with differing recommended cut scores (e.g., >= 6 vs. 7), as a reliable feigning detector in TBI (Mathias et al., 2002), chronic pain (Etherton et al., 2005), and attention-deficit/hyperactivity disorder (ADHD) samples (Marshall et al. 2010). Regarding its use in TBI, results from Mathias et al.'s (2002) study, which adhered to MND criteria from Slick et al. (1999), showed the exceptional positive predictive power of RDS in classifying MND in TBI samples, which is a finding echoed in other research on this topic (Breting & Sweet, 2013).

Another relevant EVM with some proven utility in TBI samples also makes use of the WAIS DS subtest. This strategy draws from observations made by Mittenberg et al. (1995) that individuals who feign cognitive impairment often attempt to suppress DS performance, as DS is

commonly mistaken as a task of pure memory, which is the most frequently feigned neurocognitive domain (Slick, Sherman, & Iverson, 1999). In contrast, the Vocabulary (V) WAIS subtest is rarely suppressed by feigning individuals. Therefore, a common atypical performance pattern displayed by these individuals is an unusually large discrepancy between V and DS obtained scores; however, only modest MND predictive power for TBI samples has been demonstrated, thus limiting this strategy's usage to a more supplementary role in MND detection in this population (Mittenberg et al., 1995).

While other EVMs and SVMs researched have shown modest predictive power for feigned impairment in TBI groups (e.g., DS forward, DS reverse, Trail-making test [TMT] floor-effect, Finger Tapping Test), there remains a remarkable need for the development of new strategies and instruments with better predictive reliability. As threats of coaching and test security breaches grow in concert with forensic neuropsychological evaluations, advancements in feigned neurocognitive impairment detection science are critical. This is particularly true in relation to "high stakes" forensic cases of mTBI, where the combination of elusive definitional parameters and acutely elevated malingering risks present a daunting challenge for diagnostic accuracy. One area where more feigning research might offer promise in tackling this challenge is that of sports-related brain injuries (Breting & Sweet, 2013). Specifically, the high prevalence rates of TBIs and concussions sustained by athletes in the National Football League (NFL), in conjunction with the recent finalization of the NFL Players' Concussion Injury Litigation Class Action Settlement Program, make the group of Settlement Class Members a population worthy of new research efforts.

Repeated Head Injuries in Professional Football (USA Style) Players

Compared to the general population, professional football players (USA style) are

between five and 19 times more likely to develop some form of dementia in their lifetime, according to McGrath (2011). Although this elevated probability cannot definitively be attributed solely to the increased rates of TBI in the sport, a plethora of empirical research (e.g., Gavett et al., 2011) links repeated head injuries to the long-term cognitive impairment emblematic of neurodegenerative disorders, such as dementia. Unfortunately, football players are not only subjected to an increased risk of experiencing TBI, but also of experiencing repeated TBIs (Omalu et al., 2010). This is highly problematic due to the cumulative effects of sustaining multiple head injuries. First, the risk of sustaining future head injuries appears to grow exponentially with each TBI suffered (Gualtieri & Cox, 1991). Additionally, successive TBIs, even at mild or sub-concussive levels, can result in more damage than TBIs of the same severity occurring in isolation for the first time (Omalu et al., 2010). This increased vulnerability is due, in part, to the permanent alterations in neuronal cellular functions (e.g., amplified inflammatory response) that can occur even from mild forms of TBI (Gennarelli & Graham, 2005). Despite advancements in helmet technology integrated into professional football over time, the crux of the problem remains the trauma from intracranial collisions of the brain and skull, which helmets do little to alleviate (Lezak, 2012).

Of the different types of closed head injuries that can occur when playing football, the most common is concussion, which falls on the mild end of the TBI continuum. The milder symptomology surrounding concussions contributes to the haziness in its clinical presentation, which is reflected in the absence of a consensus definition of concussion, or mTBI, in the literature (Evans, 2010). Generally, the most accepted guidelines for describing and diagnosing concussion and mTBI are those set forth by American Congress of Rehabilitation Medicine (ACRM), the National Academy of Neuropsychology (NAN), and the Third International

Conference on Concussion in Sport (ICCS) (Lezak, 2012). The latter of these definitional frameworks was purported by the ICCS to be crafted with injured athletes in mind (McCrory et al., 2009). Nonetheless, the essential elements shared across these sets of diagnostic criteria include: temporary physiological disruption of brain function induced by traumatic forces, rapid onset of impairment, and eventual spontaneous resolution of physical, cognitive, emotional, or sleep-related symptoms. The most common points of contention usually surround questions of timing and presence of loss of consciousness, posttraumatic amnesia, and neuroimaging abnormalities (Breting & Sweet, 2013).

Given the nebulous nature of the current state of mTBI criteria, it is understandable how controversy frequently engulfs this field of research. Added ambiguity stems from what is termed postconcussion syndrome (PCS). This refers to the continuation (i.e., more than three months following mTBI) or worsening (i.e., in hours to days since mTBI) of physiological, emotional, or cognitive symptoms after incurring mTBI (Evans, 2010). Two versions of PCS diagnostic criteria published by the DSM-IV and International Classification of Diseases, Tenth Edition (ICD-10), respectively, were found to differ significantly in their ultimate classifications of PCS (Boake et al., 2004). Sizeable discrepancies in diagnostic rates of the "same" condition are alarming and indicative of the pressing need for more research capable of injecting clarity into this body of literature. While literature on differences in PCS diagnostic trends between the latest iterations of these classification systems is largely nonexistent, updated guidelines in the DSM-5 call for PCS to be given a diagnosis of neurocognitive disorder (major or mild) due to TBI (American Psychiatric Association, 2013), while parallel recommendations in ICD-11 are not explicitly presented.

One last oft-debated condition worth mentioning that is highly relevant to professional

football is chronic traumatic encephalopathy (CTE). This condition is manifested through the gradually deleterious effects of repeated concussions or sub-concussive head injuries that accumulate over time (Omalu et al., 2010). CTE is a reflection of the long-term damage possible from the aforementioned deformations in cellular structure and disruptions in axonal pathways caused by externally applied mechanical forces (Chen, Smith, & Meaney, 2009). As CTE is relatively new to TBI literature, more physiological, neuropsychological, and genetic investigations are needed to illuminate the most significant risk factors and pathological underpinnings of CTE. Nevertheless, some recent findings, like those from Mez et al. (2017), which reported that 110/111 brains (99%) examined from deceased former NFL players were consistent with CTE neuropathology, while likely confounded by clustered data (i.e., brains might have been donated due to suspicion of CTE), offer some compelling evidence linking this condition to the long-term effects of playing professional football.

NFL Baseline Assessment Program

The resounding volume of research supporting a causal connection between long-term neurological problems and repeated head trauma was impetus for a class action lawsuit filed by retired NFL football players (plaintiffs) against the National Football League. The plaintiffs accused the NFL of having known about the irrefutable risks correlated with repeated TBIs and subsequently failing to alert players of these risks, while intentionally withholding relevant information from them, which the NFL denied (Legg, 2015). A settlement agreement between the plaintiffs and the NFL was negotiated and eventually finalized on January 7th, 2017 ("NFL concussion settlement," 2018). Included in the benefits of this settlement are possible monetary awards for retired players meeting criteria (as defined by settlement agreement terms) for: amyotrophic lateral sclerosis (ALS), Parkinson's disease, Alzheimer's disease, level 2

neurocognitive impairment, or level 1.5 neurocognitive impairment ("NFL concussion settlement," 2018). Also integrated into the settlement package is the Baseline Assessment Program (BAP), which entitles claimants to one free independently administered neurological and neuropsychological (baseline) assessment, the results of which help determine whether a given retired NFL player receives one of the aforementioned diagnoses, and by extension, a monetary award. Those retirees who qualify for moderate cognitive impairment are also entitled to follow-up testing, treatment, and counseling ("NFL concussion settlement," 2018).

In order to qualify for BAP monetary awards, retired NFL players must exhibit sufficient evidence of impairment, via BAP results, per diagnostic criteria set forth within the concussion settlement agreement. Specifically, for a diagnosis of level 1.5 neurocognitive impairment, there must exist: (a) concern about severe decline in retiree's cognitive function, (b) evidence from BAP neuropsychological performance of moderately to severely declined cognitive functioning, from premorbid levels, in at least one domain of executive function, learning and memory, or complex attention plus one other of these domains or from at least one domain of language or perceptual-spatial abilities, (c) a Clinical Dementia Rating (CDR) scale of 1.0 (mild) in community affairs, home and hobbies, and personal care subscales, with collateral support from relevant documentation (e.g., medical records) and (d) evidence of these deficits occurring independently from alternative explanations of medication, substance abuse, or delirium ("NFL concussion settlement," 2018). With respect to criterion (c), in the absence of corroborating documentation, functioning decline in criterion (b) must be found in at least executive function or learning and memory domains plus one other cognitive domain listed, and the addition of a third-party sworn affidavit (from a non-family member individual familiar with the retiree) endorsing the retiree's functional deficits is also required ("NFL concussion settlement," 2018).

Requirements for level 2 neurocognitive impairment parallel those of level 1.5 criteria, but criterion (b) must reflect severe decline in cognitive functioning, and criterion (c) must include a CDR scale of 2.0 (moderate). Regarding the agreed upon criteria for alternative aforementioned diagnoses qualifying a retiree for monetary benefits, relevant details are beyond the scope the current review and, thus, will not be described here.

The values of possible monetary rewards vary by type of qualifying diagnosis and age at time of qualifying diagnosis ("NFL concussion settlement," monetary award grid, 2018). Moreover, award values determined by these two variables represent monetary starting points, with reductions in value possible due to the presence of one or more extenuating conditions (e.g., 10% to 97.5% reduction depending on number of seasons played under five; 75% reduction if diagnosed with Stroke prior to date of qualifying diagnosis, etc.). Conspicuously, one factor classified as a possible extenuating condition is "a medically diagnosed TBI occurring before the qualifying diagnosis," which can result in a 75% reduction for "certain retired NFL football players" ("NFL concussion settlement," monetary award grid, 2018). While the actual concussion settlement language includes the critical detail that TBIs possible of resulting in this reduction are "unrelated to NFL football," the burden of proof lies with the retired player, as the 75% reduction is applied in such cases unless the player "can show that the TBI is not related to the qualifying diagnosis" ("NFL concussion settlement," 2018). Importantly, TBI unrelated to NFL football play is defined in the settlement as having occurred during or after a player's career and characterized by loss of consciousness (LOC) for more than 24 hours. Considering the well-documented problems associated with retrospective self-report and precise measurement of LOC, generally, the inclusion of this extenuating condition and its potential ramifications are disconcerting. Nevertheless, with possible reductions notwithstanding, monetary award values

for qualifying conditions of interest in the current study (i.e., level 1.5 and 2 neurocognitive impairment) range from $3,062,100 (level 2, under age 45) to $25,518 (level 1, age 80+) ("NFL concussion settlement," monetary award grid, 2018).

With the opportunity for receiving these substantial sums of money, the neuropsychological assessments in the NFL BAP easily qualify as "high stakes" TBI forensic evaluations, where elevated risk for feigned neurocognitive impairment is expected (Bianchini, Curtis, & Greve, 2006). The diagnostic challenges brought about by increased malingering risk are then heightened in this case by the high profile nature of the NFL concussion settlement program, which undoubtedly engenders additional validity threats from attorney coaching. Furthermore, the retrospective nature of crucial data collection, important for diagnostic determinations in these evaluations, adds to the challenge of drawing accurate conclusions. In light of this blend of unique circumstances, the NFL BAP environment offers an excellent research opportunity for answering the call for advancing the science of existing strategies for detecting feigned neurocognitive impairment, as well as for proposing new detection methods. Regarding the latter research opportunity, the prospect of integrating data from collateral reports of neurocognitive impairment (i.e., from spouse, immediate family member, close friend, etc.) into MND determination procedures is ripe with scientific potential and limited in precedent.

The Need for and Problem with Using Multiple Validity Measures

Detection of feigned neurocognitive impairment is improved by employing multiple Performance Validity Tests (PVTs) and Symptom Validity Tests (SVTs) (Victor et al., 2009). In fact, neuropsychologists self-identified as experts in validity research use eight combined stand-alone and embedded PVTs per forensic evaluation, on average (Schroeder, Martin, & Odland, 2016). However, the deployment of several validity measures on a given assessment gives rise to

multiplicity problem (Morgan, 2016), which essentially spotlights the erroneous assumption that separate tests of validity are statistically independent. If this assumption were completely true, then a clinician could multiply the first measure's pretest odds by its likelihood ratio (LR) to compute posttest odds, which could then be used as the pretest odds for the second measure, and so on. For example, if the prior probability (i.e., base rate) of malingering prior to administration of the first validity test was estimated to be 40 percent, which is converted to pretest odds via the formula Odds = Probability/(1-Probability) for a value of 0.67 [.40/(1-.40)], and the patient fails the first validity test of 0.90 specificity and 0.50 sensitivity, then the LR, which is equal to sensitivity/(1-specificity) (i.e., 5 in this case), multiplied by the pretest odds of 0.67 would yield posttest odds of 3.35 or a 77 percent posttest probability of malingering. Continuing with this example, now the base rate of malingering going into the second validity measure is no longer 40 percent, but now 77 percent, given the failure of one validity test already. Thus, a failure on the second validity test of identical specificity and sensitivity would yield posttest probability of nearly 95 percent that a diagnosis of malingering is appropriate. Unfortunately, the inherent degree of multicollinearity between most measures renders a "chaining of likelihood ratios" approach untenable due to the inflated risk of false positives.

The goal of minimizing false positives when diagnosing malingering is paramount. Misdiagnosing an individual as malingering effectively attaches to his/her history a mark of having purposefully exaggerated cognitive dysfunction to obtain an external reward or avoid formal duty, which can have devastatingly negative, long-lasting impacts on that individual's reputation, career opportunities, and personal life. Therefore, it is imperative that clinicians exercise great diligence throughout each step of the diagnostic process, remaining vigilant of psychometric pitfalls and always interpreting test results in conjunction with the patient's history

and the context of the evaluation. In relation to a test's specific psychometric attributes, a secondary consequence of failing to minimize false positives is that increasing the false positive rate of a measure weakens its ability to accurately detect invalid performance when performance is actually invalid. In the development of new or improved validity measures and diagnostic test paradigms, maintaining high specificity should never be sacrificed for the improvement of sensitivity.

The generally accepted rule of thumb stressing that PVTs and SVTs preserve at least 90 percent specificity can be easily remembered by the adage "it is better that ten guilty persons escape than that one innocent person suffer," otherwise known as the Blackstone ratio in criminal law (Blackstone, 1844). As any one test's sensitivity consequently falls with an increase in its specificity, the use of a single validity measure to detect feigned neurocognitive impairment is ultimately insufficient. Conversely, the use of multiple validity measures helps improve overall detection accuracy (Larrabee, 2003). Still, the aforementioned approach of chaining likelihood ratios across validity measures obscures the true overall rate of correct classification when combining multiple measures.

In addition to this multiplicity problem, there is the separate issue of which specific combination of validity tests is optimal for obtaining desirable overall accuracy rates when diagnosing malingering. Further complicating this question is the throng of validity measures from which to choose as well as the assortment of flexible and fixed batteries employed across and within various clinician subspecialty groups and client populations (Millis, 2001). To help elucidate a solution to the entangled problems of multiplicity and variable test collections, Larrabee (2019) sought to demonstrate that aggregating failures across multiple PVTs/SVTs as a method of discerning valid from invalid performers is largely equivalent in overall classification

accuracy as a logistic regression in which *all* the PVTs and SVTs from the study were used as continuous variables. By showing that a paradigm of tabulating failures on across multiple PVTs/SVTs (at acceptable per-test false positive rates), in which the number of failed tests represents a criterion cutoff, reliably differentiates valid/invalid group membership, clinicians utilizing variable or flexible test batteries can better depend on the accuracy of failing of multiple validity tests as a marker of poor effort, *regardless of the specific combination of PVTs/SVTs*.

While the question of which specific PVTs/SVTs are "best" remains a valid and important question still warranting attention in practice and meriting future research, Larrabee's (2019) study alleviates some of the burden in necessitating the same validity tests or combination of tests be used for every evaluation in which the threat of malingering is prominent. A different question more suitably answered by studies on diagnostic paradigms of multiple validity measures is how many PVTs/SVTs is the optimal amount. Results from Larrabee (2019) suggested a criterion of three or more failures on tests with satisfactory specificity (i.e., approximately 90%) yielded the best overall classification accuracy, while Davis and Millis (2014) indicated two or more failures as an acceptable cutoff score.

In summary, there is strong evidence supporting the need to utilize multiple validity measures in neuropsychological evaluations (Heilbronner et al., 2009), as well as lingering limitations and unanswered questions revealing the problems with doing so. If every neuropsychological evaluation had the exact same battery, then over time a logistic regression formula could be derived which would assign reliable weights to each measure reflecting its differential importance and relevance for predicting suboptimal effort. However, this is obviously not the case, as most clinical neuropsychologists use flexible or varied batteries and diverge in their approaches to test selection. Therefore, there is great clinical value in replicating

and expanding research demonstrating that one can aggregate the number of tests failed across any combination of validated PVTs/SVTs and then use that number of failures as a criterion score for accurately detecting valid or invalid performance. One fitting way to do build a suitable study with this aim is with a large sample in which all subjects complete every test within a fixed battery, in the presence of a clear external incentive (to meet Slick criteria), and are classified as malingering or not malingering based on their performance on more than one PVT. Then, the remaining measures in the battery would be employed as embedded PVTs/SVTs and used as independent continuous variables in a logistic regression to predict probability of poor effort. Finally, if a diagnostic approach based on aggregating the number of failures on embedded PVTs/SVTs at the accepted specificity rate of 90% is shown to have similar overall prediction accuracy as the logistic regression formula, it would further support the approach of diagnosing poor effort via examining the number of failed validity measures across a variety of test batteries.

Proposed Study

The proposed study seeks to examine the rates of feigned neurocognitive impairment within a sample of retired NFL football players who completed standardized neuropsychological assessments as part of the NFL BAP and evaluate the classification accuracy of an aggregated validity test failure paradigm. First, archival data obtained from retirees will be carefully inspected, with variables of interest including quantitative neuropsychological test data required for the determination of possible, probable, or definite MND, outlined by Slick et al. (1999). Specifically, obtained scores on the TOMM, ACS Word Choice, MSVT, and Reliable Digit Span (RDS) will be used to categorize players into Valid and Invalid groups, with one below chance failure on any of the three forced-choice measures (i.e., not RDS) or two or more failures on the four validity measures used as Invalid group inclusion criteria. Second, a parsimonious model for

group prediction will be created from multiple cognitive tests and MMPI-2-RF response bias indices selected to serve as embedded PVTs/SVTs based on theory, support from the literature on effort and response bias detection, and classification statistics. Test selection will also emphasize diversity in the neuropsychological constructs assessed and test formats. The model will utilize a group membership classification paradigm in which the number of PVT/SVT failures, with PVT/SVT pass/fail cut scores set at a per test false positive rate of 10 percent for the sample, serves as the criterion cutoff. Third, a logistic regression will be performed with scores on all PVTs/SVTs selected for classification analysis used as continuous independent variables predicting group membership. Fourth, the predictive accuracy of the different failure tabulation paradigms (e.g., criterion of two or more failures, three or more failures, etc.) will be compared to the predictive accuracy of the logistic regression formulas for PVT/SVT, PVT-only, and SVT-only combinations. Lastly, individual PVTs/SVTs will be compared to each other and to expectations from previous studies in relation to their respective classification accuracy statistics.

Hypotheses

1. Findings from the proposed study will be similar to Larrabee (2019) and show that a combination of three or more failures on cognitive and/or personality measures utilized as PVTs/SVTs with per test false positive rates of 10 percent will yield diagnostic accuracy comparable to a logistic regression formula in which all cognitive and personality scores are used as continuous variables to predict Valid/Invalid group membership.

2. The BDAE-CIM and WMS-IV Recognition subtests will have the best predictive accuracy among all cognitive tests and personality indices utilized as PVTs/SVTs, in accordance with expectations based on evidence in the literature supporting their use as embedded forced-choice validity measures (e.g., Erdodi, 2016; Holdnack, 2013).

3. Rates of MND classification in the sample population of retired NFL players studied will be comparable to 38%, aligned with expectations for malingering in "high stakes" neuropsychological evaluations involving litigating mild TBI samples, as shown in Mittenberg, Patton, Canyock, and Condit (2002).

CHAPTER 2

METHOD

Participants

The study utilized archival data from 265 adult males who retired from NFL prior to participating in the BAP assessment. All subjects in the current study met the primary inclusion criteria of having completed at least one BAP fixed-battery neuropsychological evaluation and being age 69 or younger at the time of the evaluation. Retired players aged 70 and older were excluded from the study for two primary reasons: First, as outlined by the BAP guidelines, the fixed-battery for players 70 and older differed from that of the younger group for several measures across multiple domains of cognitive functioning. Thus, direct comparisons on several tests of interest were not possible. Second, given the higher risk of neurodegenerative disorders and other medical conditions with potentially deleterious effects on cognition (e.g., diabetes mellitus type II, cardiovascular disease, etc.) associated with advanced age, exclusion of this older demographic helps mitigate the confounding impact of such factors on analytic inferences. Notably, as Dean et al. (2009) showed, within a heterogeneous sample of individuals with dementia in the absence of an external incentive, traditional cutoff scores on commonly employed effort indices exhibited inflated false positive rates, suggesting that the application of such cutoffs in research samples confounded by high risks of dementia obscures interpretive clarity.

Demographic information for the entire sample is depicted in Table 1. Mean age for the entire sample was 49.1 (range = 28 – 69). Average number of seasons played was 5.3 (range = 1 – 15). Years of education ranged from 14 to 20, with an average of 16.0. For the entire sample, 220 players (83%) were of African-American descent and 45 players (17%) were of Caucasian

descent.

Table 1

Demographic Means and Standard Deviations for All Cases, Valid Group, and Invalid Group

Demographic or measure		All Cases N = 265	Valid Group n = 196	Invalid Group n = 69	*t*	*p*	*d*
Age (Years)	*M*	49.1	50.1	46.5	-2.50	.013	-0.35
	(SD)	(10.2)	(10.4)	(9.0)			
Education (Years)	*M*	16.0	16.0	15.8	-1.86	.064	-0.26
	(SD)	(1.0)	(1.0)	(0.9)			
Seasons played	*M*	5.3	5.5	4.9	-1.27	.204	-0.18
	(SD)	(3.1)	(3.2)	(3.0)			
TOPF-WR score	*M*	36.2	37.9	31.2	-3.76	<.001	-0.51
	(SD)	(13.0)	(12.7)	(12.6)			

Inclusion Criteria for Validity Group Determination

The 265 players in the study were divided into two groups (i.e., Valid and Invalid), and further classified into one of two subgroups (i.e., good effort or possible poor effort and probable poor effort or definite poor effort, respectively) based on their scores on four measures of performance validity. The following formulaic sequence, which parallels the Slick criteria, was utilized for group determination: 1) If a below chance score was obtained on TOMM Trial 2 (i.e., <25), ACS Word Choice (i.e., <25), or on immediate, delayed, or consistency trials of the MSVT (i.e., <50%), a designation of definite poor effort (invalid group) was assigned. 2) For the remaining players, if two of the four well-validated performance validity measures were failed (i.e., scoring at or below recommended cutoff scores), a designation of probable poor effort (invalid group) was assigned. 3) The remaining players were assigned to the valid group, with a designation of possible poor effort given for one performance validity failure and a designation

of good effort given for zero failures on the four performance validity measures.

After applying the above formula, 196 players (74.0%) were classified as valid performers (150 [56.6%] = good effort; 46 [17.4%] = possible poor effort) and 69 players (26.0%) were classified as invalid performers (53 [20.0%] = probable poor effort; 16 [6.0%] = definite poor effort). Comparison demographic information is presented in Table 1. Education and number of seasons played did not differ significantly between groups. However, average age of players assigned to the valid group (m = 50.1, SD = 10.4) was significantly higher than average age of players in the invalid group (m = 46.5, SD = 9.0), albeit with a relatively low effect size (d = -0.35). While scores on a reading task utilized for premorbid estimates of intellectual functioning (TOPF-WR) also differed significantly between groups, the interpretive value of this results is limited due to its technical designation as a performance measure, thereby rendering it vulnerable to feigned impairment.

Measurement Approaches

Quantitative test data (i.e., data recorded from standardized psychometric instruments) from BAP assessment fixed-batteries was gathered systematically and fully de-identified for analysis in the current study.

Premorbid Intellectual Functioning Test: TOPF-WR

The Test of Premorbid Functioning – Word Reading (TOPF-WR) is a revision of the Wechsler Test of Adult Reading (WTAR), a stand-alone reading test used to estimate premorbid intellectual and memory abilities. The TOPF-WR, like the WTAR, asks the examinee to pronounce increasingly difficult words with atypical grapheme-to-phoneme translation, which maximizes assessment of previous learning instead of current ability to apply pronunciation rules (Nelson & Willison, 1991). Reading recognition has been shown to be relatively resistant to

cognitive decline, and thus a good estimate of FSIQ. Estimating premorbid functioning is useful for establishing a comparison to determine if a significant decline in ability has occurred, especially in the context of suspected loss within incentivized settings.

Cognitive Tests

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

The WAIS-IV is the most widely used measure of intelligence in the U.S. (Hartman, 2009). Ten subtests from the WAIS-IV were administered in the BAP battery. Three subtests (i.e., Digit Span [DS], Arithmetic [AR], and Letter-Number Sequencing [LNS]) were from the Working Memory Index, designed to assess ability to temporarily hold and manipulate auditory information in mind. Three subtests (i.e., Symbol Search [SS], Coding [CD], and Cancellation [CA]) were from the Processing Speed Index, designed to assess speed of mental processing and graphomotor skills. Three subtests (i.e., Block Design [BD], Matrix Reasoning [MR], and Visual Puzzles [VP]) were from the Perceptual Reasoning Index, designed to assess fluid reasoning, nonverbal problem solving, and pattern recognition. Similarities (SI), designed to assess abstract verbal reasoning, was the only subtest used from the Verbal Comprehension Index.

Phonemic and Semantic Verbal Fluency

The Controlled Oral Word Association Test (COWAT) is a measure of phonemic verbal fluency in which the examinee is given a letter of the alphabet and instructed to name as many different, non-proper nouns as possible in one minute. This process is repeated two more times, each time with a different letter, with the total raw score equating to the sum of correct words spontaneously produced over the three trials. Semantic (or category) verbal fluency was measured with the Animal Naming Test, in which the examinee is given the category (i.e., animals) and instructed to name as many different subtypes of the category as possible in one

minute.

Trail Making Test (Trails A & B)

A measure of focused visual attention and graphomotor speed, Trails A is a numerical sequencing task which was administered within the current battery to serve primarily as a foundation for the introduction of Trails B. Trails B is a more challenging alphanumeric switching task, which taps motor speed, processing speed, as well as mental flexibility and inhibitory control.

Booklet Category Test (BCT)

The BCT is a measure of executive cognitive functioning on which good performance is associated with intact pattern recognition, effective discernment of systematic problem-solving strategies, rule learning, and nonverbal concept formation.

Wechsler Memory Scale, Fourth Edition (WMS-IV)

Three subtests from the WMS-IV were utilized as primary measures of immediate memory, delayed memory, and recognition memory. The Logical Memory I (LM I) subtest assesses the examinee's ability to listen and repeat back auditory information presented within a narrative context (i.e., stories). After a 20-30 minute delay following administration of LM I, the Logical Memory II (LM II) subtest measures the ability to accurately recall story information from LM I. With no feedback given, the examiner then administers Logical Memory Recognition (LM Recog), which consists of 30 yes/no questions regarding information from the original stories. Immediate, delayed, and recognition memory for a list of word pairs is similarly assessed via the sequential administration of Verbal Paired Associates I (VPA I), II (VPA II), and Recognition (VPA Recog). Immediate visual memory for increasingly complex designs is

assessed with the Visual Reproduction I (VR I) subtest, in which the examinee is presented an abstract line drawing for 10 seconds before being instructed to draw the figure from memory immediately after the stimulus is removed. Visual Reproduction II (VR II) is administered 20-30 minutes after VR I with the examinee asked to reproduce the designs from VR I without cues. Finally, on the Visual Reproduction Recognition (VR Recog) trial, the examinee is instructed to identify the correct figures from multiple -choice options. Notably, the use of WMS-IV recognition trials as embedded performance validity indicators has been proposed in literature on feigned neurocognitive impairment (Holdnack et al., 2013).

Boston Naming Test (BNT)

The BNT is a language-related measure of visual confrontation naming. The examinee is presented with and asked to name line drawings of objects ranging from common to less commonly known. Most healthy adults tend to score highly on the BNT, while low scores can occur in a range of clinical conditions (e.g., Alzheimer's disease, temporal lobe epilepsy, left-hemisphere cerebrovascular accidents, etc.; Henry et al., 2004; Randolph et al., 1999; Kohn & Goodglass, 1985) or in the presence of feigned neurocognitive impairment.

Boston Diagnostic Aphasia Examination – Complex Ideational Material (BDAE-CIM)

The BDAE-CIM is a 12-item sentence comprehension task designed to assess receptive language. However, patients with apparently intact language functioning often perform in the impaired range on this assessment suggesting its potential use as a Performance Validity Test (Erdodi et al., 2016).

Personality and Psychopathology Test: MMPI-2-RF

The Minnesota Multiphasic Personality Inventory, Second Edition, Restructured Form

(MMPI-2-RF) is a self-report measure of personality and psychopathology derived from the MMPI-2. It was designed to for improved administration efficiency and enhanced construct validity, as it is shorter than the MMPI-2 with 338 of the most clinically significant items retained. For the scope of the current study, only validity scales were analyzed; namely, F-r (infrequent responses in the general population), Fp-r (infrequent responses in a psychiatric population), FBS-r (level of somatic/cognitive complaints associated with over-reporting), and RBS (exaggerated memory complaints).

Performance Validity Tests

Test of Memory Malingering (TOMM)

The TOMM is one of the most commonly used and well-validated stand-alone measures of performance validity (Sharland & Gfeller, 2007). It consists of two learning trials in which the examinee is presented with 50 line drawings of common, everyday objects. Each learning trial is followed by a forced-choice recognition trial. Importantly, the examiner delivers verbal feedback as to whether the examinee's responses on the recognition trials are correct or incorrect. Depending on the score obtained on Trial 2 (i.e., if the examinee does not score above the recommended cutoff), an optional Retention Trial of recognition can be administered 15 minutes after Trial 2. With adequate effort, a cutoff score of 45/50 or higher on Trial 2 or Retention Trial has been shown to be reliably obtained by patients across a wide range of clinical samples (e.g., TBI [Bauer et al., 2007], pain-related disability [Greve et al., 2009], learning disabilities [Lindstrom et al., 2009], etc.). That is, there is robust support in the literature that the TOMM possesses high sensitivity and specificity for detecting poor effort, regardless of underlying psychiatric or neurologic conditions.

ACS Word Choice (WC)

The WC subtest of the ACS package is a stand-alone, forced choice performance validity measure. Examinees are first directed to view common words one at a time and identify whether the words are man-made or natural, as a method of focusing their attention to the stimuli during the learning trial. Then the examiner presents a list of 50 word pairs, with each pair containing one word from the previous learning trial. The examinee is instructed to choose the correct word from each pair. In the ACS validation study (Holdnack & Drozdick, 2009), an overall clinical sample comprised of individuals with neurologic, psychiatric, and developmental disorders, obtained a score of 43/50 or higher on WC at a base rate of 10 percent or less. This cutoff was used in the Valid/Invalid Group determination process for the current study.

Medical Symptom Validity Test (MSVT)

The MSVT is a forced-choice, stand-alone performance validity measure in which the examinee is presented with a list of 10 semantically related word pairs on two learning trials. Following the learning trials, immediate and delayed (i.e., 10 minutes) forced-choice recognition trials are administered, which together with a consistency measurement of responses on the two recognition trials comprises the MSVT effort measurements. The MSVT manual (Green, 2004) validation sample exhibited a ceiling effect on immediate recognition, delayed recognition, and consistency subtests. The suggested cutoff score of 85 percent or less correct on any of the three indices demonstrated adequate sensitivity and specificity, as scores below this cutoff reflected performance two standard deviations below the mean for consistently responding individuals in the normative sample.

Reliable Digit Span (RDS)

Derived from the WAIS-IV DS subtest, RDS is a well-researched embedded validity

measure representing the sum of the longest span of digits correctly recited on both trials of an item for both DS Forward and DS Backward subtests. Numerous studies suggest a cutoff score of seven or less is adequately sensitive for reliably detecting feigned neurocognitive impairment in samples of individuals with TBI (Mathias et al., 2002), toxic exposure (Greve et al., 2007), and pain-related disorders (Etherton et al., 2005). However, given some evidence in the literature of unacceptably high false positive rates at this cutoff (e.g., Babikian et al., 2006), a more conservative cutoff of six or less was utilized for Valid/Invalid Group designation the current study.

Procedure

Archival data was pulled from the files of retired NFL players who successfully completed neuropsychological assessments within the broader framework of the NFL BAP. As formalized informed consent to take part in BAP processes was necessarily obtained from individuals who had concluded their BAP examinations prior to their potential inclusion in the current study, follow-up consent for use of their de-identified data in the proposed analyses is considered supplementary. The current study examined differences in scores obtained on standardized measures of cognitive ability, performance validity, and symptom validity across Valid and Invalid groups, defined in accordance with Slick criteria.

CHAPTER 3

RESULTS

An independent-samples t-test was conducted to compare scores on all cognitive performance measures and MMPI-2-RF symptom validity indices between valid and invalid groups. Raw scores were utilized for cognitive measure analyses for two primary reasons. First, the absence of a statistical conversion to t-scores allows for more direct comparisons unburdened by the data loss that can occur with transformative procedures. That is, the same raw score obtained by a 49-year-old and a 50-year-old on a given test, reflecting a broadly equivalent finding, suffers informational distortion when subjected to a t-score conversion based on tiered age groupings of 40-49 year-olds and 50-59 year-olds, which would falsely dichotomize otherwise minimally different scores. Second, evaluators exercised clinical judgment in deciding whether full demographic adjustments or education-only adjustments were utilized, which further adds confounding variance into interpretability of t-scores. Conversely, t-scores for MMPI-2-RF data were used for comparisons in the current study, as demographic information is not involved in the computation of these t-scores, thereby liberating them from the confounding issues encumbering cognitive t-scores. Table 2 displays the means, standard deviations, and effect sizes for all cognitive test scores and embedded symptom validity scales for the entire study sample, as well as for valid and invalid groups specifically. There were significant differences between groups on all cognitive scores at a probability level below 0.0001, except for Block Design (p = .012). Similarly, all symptom validity scores were significantly discrepant between groups, with probabilities below 0.0001 for three of the four MMPI-2-RF indices (FP-r p = 0.001). Effect sizes (absolute value) for cognitive measures ranged from d = 0.35 to d =1.26, with the three largest effect sizes observed at d > 1.00 for VR Recognition, BDAE-CIM, and

VPA Recognition (-1.26, -1.17, and -1.08, respectively). Effect sizes for symptom validity scores were smaller on average, ranging from d = 0.63 (FP-r) to d = 0.89 (RBS).

Another descriptive statistic calculated to help select the best potential cognitive measures for further analysis was skewness (Larrabee et al., 2019), displayed in Table 3. Notably, mean absolute value of skewness for three of the four validity measures utilized in accordance with Slick criteria for initial player group classification was larger for the valid group (m = 2.77, SD = 1.53) than for the invalid group (m = 1.23, SD = 0.69), which was expected (note: while MSVT pass/fail information was available for all players in the study, not all raw MSVT scores were accessible in the archival dataset, and, thus, skewness was not calculated for this measure). For cognitive measures, mean absolute value of skewness was 0.50 (SD = 0.61) for the valid group and 0.52 (SD = 0.36) for the invalid group. However, for the seven cognitive measures ultimately selected for further analysis, mean absolute value of skewness was larger for the valid group (m = 1.08, SD = 0.92) than for the invalid group (m = 0.561, SD = 0.37).

Receiver Operating Characteristic (ROC) curve analysis also provided valuable information for assessing of the diagnostic usefulness of cognitive and symptom validity measures. In ROC, Area Under the Curve (AUC) statistics between 0.7 and 0.8 are generally considered acceptable (Mandrekar, 2010), while the 0.8 to 0.9 range is considered excellent and values over 0.9 are outstanding. AUC statistics for all cognitive and MMPI-2-RF scores were significant at a level less than 0.001 (with the exception of Block Design [p = 0.003]), ranging from AUC = 0.620 (Block Design) to AUC = 0.820 (BDAE-CIM) for cognitive measures and from AUC = 0.647 (FPr) to AUC = 0.753 (RBS). Mean AUC for the 11 total measures used for further analysis was 0.750. Table 4 shows full ROC statistics for chosen measures.

Table 2

Means and Standard Deviations for Cognitive Tests and Embedded MMPI-2-RF Symptom Validity Indices for All Cases, Valid Group, and Invalid Group

Measure		All Cases N = 265	Valid Group n = 196	Invalid Group n = 69	*t, U*	*p*	*d*
DS	*M (SD)*	24.22 (5.74)	25.69 (5.34)	20.03 (4.73)	-7.80	<.0001	-0.99
AR	*M (SD)*	12.48 (3.34)	13.17 (3.20)	10.52 (2.92)	-6.05	<.0001	-0.79
LNS	*M (SD)*	17.04 (3.38)	17.78 (2.93)	14.94 (3.71)	-6.43	<.0001	-0.84
CD	*M (SD)*	50.20 (14.60)	53.39 (13.39)	41.13 (13.98)	-6.46	<.0001	-0.84
SS	*M (SD)*	24.28 (7.93)	25.63 (7.36)	20.45 (8.30)	-4.86	<.0001	-0.65
CA	*M (SD)*	29.86 (8.94)	31.5 (8.00)	25.22 (9.34)	-5.27	<.0001	-0.70
FAS	*M (SD)*	33.66 (10.59)	35.27 (10.50)	29.10 (9.55)	-4.29	<.0001	-0.58
Tr. B	*M (SD)*	106.90 (57.26)	93.34 (44.21)	145.42 (71.34)	3501.0	<.0001	0.91
BCT	*M (SD)*	70.42 (29.93)	65.14 (29.01)	85.43 (27.50)	5.07	<.0001	0.68
SI	*M (SD)*	22.91 (5.40)	23.85 (4.84)	20.25 (6.03)	4339.0	<.0001	-0.67
LMI	*M (SD)*	18.35 (6.74)	19.71 (6.57)	14.46 (5.64)	-5.92	<.0001	-0.78
LMII	*M (SD)*	13.17 (6.83)	14.64 (6.79)	8.97 (4.97)	3458.0	<.0001	-0.83
LMRec	*M (SD)*	22.43 (3.56)	23.33 (3.14)	19.88 (3.45)	-7.64	<.0001	-0.97
VPAI	*M (SD)*	21.46 (9.92)	23.29 (9.99)	16.29 (7.67)	-5.29	<.0001	-0.71
VPAII	*M (SD)*	6.97 (3.30)	7.65 (3.18)	5.04 (2.85)	-6.02	<.0001	-0.79
VPARec	*M (SD)*	34.42 (5.37)	35.94 (4.25)	30.12 (5.91)	2487.0	<.0001	-1.08
VRI	*M (SD)*	31.49 (6.07)	32.58 (5.58)	28.41 (6.40)	-5.14	<.0001	-0.69
VRII	*M (SD)*	20.36 (9.19)	22.23 (8.86)	15.04 (8.01)	-5.94	<.0001	-0.78
VRRec	*M (SD)*	4.98 (1.77)	5.57 (1.29)	3.33 (1.91)	2404.0	<.0001	-1.26

Measure		All Cases *N* = 265	Valid Group *n* = 196	Invalid Group *n* = 69	*t, U*	*p*	*d*
BNT	*M (SD)*	49.75 (6.35)	51.08 (5.08)	45.97 (7.92)	4016.5	<.0001	-0.80
An.	*M (SD)*	17.04 (4.88)	18.05 (4.64)	14.17 (4.43)	-6.04	<.0001	-0.79
BDAE	*M (SD)*	9.80 (1.93)	10.38 (1.48)	8.13 (2.09)	2434.5	<.0001	-1.17
BD	*M (SD)*	31.43 (10.47)	32.39 (10.05)	28.72 (11.22)	-2.52	.012	-0.35
VP	*M (SD)*	11.37 (3.76)	12.09 (3.65)	9.32 (3.30)	-5.55	<.0001	-0.74
MR	*M (SD)*	14.49 (5.44)	15.6 (5.25)	11.33 (4.72)	3740.5	<.0001	-0.78
Fr*	*M (SD)*	78.60 (27.43)	72.95 (23.54)	94.62 (31.30)	3947.5	<.0001	0.79
FPr*	*M (SD)*	61.16 (20.51)	57.81 (15.21)	70.67 (29.07)	4773.0	.001	0.63
FBSr*	*M (SD)*	69.82 (14.95)	66.87 (13.93)	78.19 (14.70)	5.72	<.0001	0.76
RBS*	*M (SD)*	81.24 (19.77)	76.68 (18.12)	94.19 (18.58)	6.86	<.0001	0.89

Note. *t-scores; DS = Digit Span; AR = Arithmetic; LNS = Letter-Number Sequencing; CD = Coding; SS = Symbol Search; CA = Cancellation; FAS = Phonemic Fluency; Tr. B = Trails B; BCT = Booklet Category Test; SI = Similarities; LMI = Logical Memory I; LMII = Logical Memory II; LMRec = Logical Memory Recognition; VPA I = Verbal Paired Associates I; VPA II = Verbal Paired Associates II; VPARec = Verbal Paired Associates Recognition; VRI = Visual Reproduction I; VRII = Visual Reproduction II; VRRec = Visual Reproduction Recognition; BNT = Boston Naming Test; An. = Animal Naming; BDAE = Boston Diagnostic Aphasia Examination –Complex Ideational Material; BD = Block Design; VP = Visual Puzzles; MR = Matrix Reasoning; MMPI-2-RF Fr = Infrequent Responses; MMPI-2-RF FPr = Infrequent Psychopathology Responses; MMPI-2-RF FBSr = Fake Bad Scale; MMPI-2-RF RBS = Response Bias Scale.

Table 3

Skewness of Distributions for Independent and Dependent Variables for All Cases, Valid Group, and Invalid Group

Measure	All Cases N = 265	Valid Group n = 196	Invalid Group n = 69
TOPF-WR	0.203	0.121	0.545
RDS	1.596	1.766	2.028
TOMM 2	-2.416	-4.536	-0.853
ACS Word Choice	-2.141	-2.010	-0.809
Mean AV Skewness of PVTs	2.051	2.771	1.230
Digit Span (DS)	0.369	0.347	1.018
Arithmetic (AR)	0.441	0.379	0.985
Letter-Num. Seq. (LNS)	-0.965	-0.633	-1.176
Coding (CD)	0.022	0.093	0.258
Symbol Search (SS)	0.029	0.092	0.299
Cancellation (CA)	0.086	0.106	0.670
FAS	0.235	0.215	0.210
Trails B (Tr. B)	1.601	1.805	0.863
Booklet Category (BCT)	-0.06	0.108	-0.618
Similarities (SI)	-0.342	-0.288	0.018
Logical Memory I (LMI)	0.223	0.176	0.216
Logical Memory II (LMII)	0.393	0.256	0.389
LM Recognition (LMRec)	-0.352	-0.277	-0.173
V. P. Assoc. I (VPAI)	0.237	0.098	0.188
V. P. Assoc. II (VPAII)	-0.024	-0.135	0.144
V. P. Assoc. Rec. (VPARec)	-1.712	-2.731	-0.712
Visual Repro. I (VRI)	-0.467	-0.469	-0.222
Visual Repro. II (VRII)	-0.115	-0.252	0.093
Visual Repro. Rec. (VRRec)	-0.854	-0.859	0.107
Boston Naming Test (BNT)	-1.263	-0.731	-1.057
Animals (An.)	0.186	0.089	0.727
BDAE-CIM (BDAE)	-1.171	-1.086	-0.842
Block Design (BD)	0.478	0.439	0.775
Visual Puzzles (VP)	0.697	0.733	0.966
Matrix Reasoning (MR)	-0.046	-0.213	0.339
Mean AV Skewness of Cognitive Tests	0.495	0.504	0.523

Measure	All Cases *N* = 265	Valid Group *n* = 196	Invalid Group *n* = 69
F-r (t-score)	0.931	0.936	0.522
FP-r (t-score)	2.577	0.965	2.421
FBS-r (t-score)	0.142	0.062	0.106
RBS (t-score)	0.242	0.299	-0.046
Mean AV Skewness of MMPI-2-RF indices	0.973	0.565	0.774

Table 4

ROC Area Under the Curve (AUC) Statistics for Chosen PVTs/SVTs

PVT/SVT	AUC	Std. Error	Sig.	95% CI	
				Lower Bound	Upper Bound
Coding	0.738	0.035	<.0001	0.669	0.807
Trails B	0.741	0.034	<.0001	0.674	0.808
LMRec	0.767	0.033	<.0001	0.703	0.832
VPARec	0.816	0.030	<.0001	0.757	0.875
VRRec	0.822	0.031	<.0001	0.761	0.884
BDAE-CIM	0.820	0.028	<.0001	0.766	0.874
Visual Puzzles	0.729	0.038	<.0001	0.655	0.803
Fr	0.708	0.036	<.0001	0.637	0.779
FPr	0.647	0.039	<.0001	0.571	0.723
FBSr	0.709	0.036	<.0001	0.639	0.780
RBS	0.753	0.034	<.0001	0.687	0.819

Table 5 displays the test characteristics of interest for the seven cognitive measures and four embedded symptom validity indices selected for diagnostic prediction analysis. Cutoff scores were chosen such that each measure's false positive rate was as close to 0.10 as possible (i.e., specificity nearest to 90%). Consequently, the sensitivity statistics (Sn) shown in Table 5 reflect each measure's ability to *rule-in* invalid group membership at the corresponding specificity rates (Sp), which all hover around ninety percent. For all 11 measures in Table 5, mean Sp was 0.904 (*SD* = 0.02) and mean Sn was 0.419 (*SD* = 0.11). For cognitive measures and

MMPI-2-RF indices, mean Sp was 0.904 (*SD* = 0.02) and 0.905 (*SD* = 0.02) and mean Sn was 0.474 (*SD* = 0.90) and 0.322 (*SD* = 0.06), respectively.

Also included in Table 5 are the cognitive domains each performance measure was designed to assess, as well as the type of responses captured by each embedded symptom validity index on the MMPI-2-RF. With respect to cognitive tests, inclusion of a broad range of targeted domains in the predictive model was a choice made in an effort to theoretically minimize multicollinearity. Additionally, as the current study aimed to be generalizable so as to help inform clinical decision-making in similar populations, it was important for the model to present an array of test options reflecting some of the most common tests administered in neuropsychological evaluations. The WAIS-IV, WMS-IV, Trail Making Test, and Halstead-Reitan Neuropsychological Battery ranked first, second, third, and tenth, respectively, in a survey of most frequently used assessment instruments by clinical neuropsychologists (Rabin, Paolillo, & Barr, 2016). This provided supplementary support, on top of optimal classification statistics, cognitive domain variety, and model parsimony, for ultimately selecting Coding, Visual Puzzles, Trails B, Logical Memory Recognition, Verbal Paired Associates Recognition, Visual Reproduction Recognition, and BDAE-CIM as the cognitive measures to be used as PVTs for the current study. Notably, Digit Span was excluded from consideration, as Reliable Digit Span (RDS; one of the four well-validated effort measures used for initial group definition) is derived from two of three subtests within Digit Span (i.e., Digits Forward and Digits Backward). Finally, as the MMPI-2-RF is the most frequently used mood/personality assessment instrument by clinical neuropsychologists (Rabin, Paolillo, & Barr, 2016), possessed adequate classification statistics, and offers potentially valuable information about using SVTs in predicting effort, all four of its symptom validity indices of interest were also included in the final model.

Table 5

Test Characteristics of Interest for Chosen PVTs/SVTs

PVT/SVT	Cutoff	Spec.	Sens.	*d*	ROC AUC	Valid group skewness	Cognitive domain/MMPI-2-RF item type
Coding	≤37	0.908	0.391	-0.842	0.738	0.093	Attention/Processing Speed
Trails B	≥147	0.903	0.391	0.909	0.741	1.805	Executive Functioning
LMRec	≤19	0.888	0.449	-0.969	0.767	-0.277	Verbal Memory
VPARec	≤32	0.888	0.478	-1.083	0.816	-2.731	Verbal Memory
VRRec	≤3	0.944	0.551	-1.262	0.822	-0.859	Visual Memory
BDAE-CIM	≤8	0.872	0.507	-1.167	0.820	-1.086	Language
Visual Puzzles	≤7	0.923	0.406	-0.737	0.729	0.733	Visuospatial
Fr	≥109	0.913	0.304	0.790	0.708	0.936	Infrequent responses in general population
FPr	≥81	0.923	0.246	0.627	0.647	0.965	Infrequent responses in psychiatric population
FBSr	≥85	0.898	0.348	0.757	0.709	0.062	Over-reporting of somatic/cognitive complaints
RBS	≥99	0.888	0.536	0.886	0.753	0.299	Exaggerated memory complaints

Table 6

False Positive Rate Set to be as Close as Possible to 10% for Each PVT/SVT

PVT/SVT measures (VTs)	Cutoff	Sensitivity	False +	# FPs	OC%	*N*
Coding (raw)	≤37	0.391	0.092	18	77.4	265
Trails B (seconds)	≥147	0.391	0.097	19	77.0	265
LMRec (raw)	≤19	0.449	0.112	22	77.4	265

PVT/SVT measures (VTs)	Cutoff	Sensitivity	False +	# FPs	OC%	*N*
VPARec (raw)	≤32	0.478	0.112	22	78.1	265
VRRec (raw)	≤3	0.551	0.056	11	84.2	265
BDAE-CIM (raw)	≤8	0.507	0.128	25	77.7	265
Visual Puzzles (raw)	≤7	0.406	0.077	15	78.9	265
Fr (t-score)	≥109	0.304	0.087	17	75.5	265
FPr (t-score)	≥81	0.246	0.077	15	74.7	265
FBSr (t-score)	≥85	0.348	0.102	20	75.5	265
RBS (t-score)	≥99	0.536	0.112	22	79.6	265
Mean	---	0.419	0.096	18.7	77.8	265

	Group							
	Invalid				Valid			
	≥2	≥3	≥4	≥5	≥2	≥3	≥4	≥5
Fail VTs	58	53	43	37	47	34	23	8
Pass VTs	11	16	26	32	149	162	173	188

Cutoff	Sensitivity	False +	# of FPs	OC %	*n* Correct
≥2 VTs	0.841	0.240	47	78.1	207
≥3 VTs	0.775	0.173	34	81.3	217
≥4 VTs	0.623	0.117	23	81.5	216
≥5 VTs	0.536	0.041	8	84.9	225

# VTs Failed	Group	
	Invalid	Valid
F = 0	4	109
F = 1	7	40
F = 2	5	13
F = 3	10	11
F = 4	6	15
F = 5	16	3
F = 6	7	5
F = 7	2	0
F = 8	4	0
F = 9	5	0
F = 10	3	0
F = 11	0	0
Total	69	196

Table 6 displays group classification data in predicting whether players' evaluations were designated as Valid or Invalid for the seven PVTs (i.e., cognitive measures chosen from the battery) and four SVTs (i.e., response bias indicators embedded in the MMPI-2-RF), as a function of setting a per test false positive rate at 10 percent. Additionally, Table 6 presents group prediction statistics for using a cutoff of failure of two or more, three or more, four or more, and five or more PVTs/SVTs. Raw numbers of group membership are also shown when applying cutoff criteria of zero failures to cutoff criteria of failure on all 11 validity tests (VTs). Failure of two or more PVTs/SVTs had a Sn equal to 0.841 and a Sp equal to 0.760. When the cutoff criterion was changed to failure of three or more PVTs/SVTs, Sn was equal to 0.775 and Sp was equal to 0.827. Failure of four or more PVTs/SVTs resulted in a Sn decrease to 0.623 and a Sp increase to 0.883 (the closest to 90 percent Sp of all of the cutoff criteria in the aggregated PVT/SVT failure paradigm). Changing to the criterion to failure of five or more PVTs/SVTs increased Sp to 0.959 but decreased Sn to 0.536). The top three overall rates of correct prediction (OC%) across all PVT/SVT failure cutoffs were 81.3 percent, 81.5 percent, and 84.9 percent for three, four, and five PVT/SVT failures, respectively.

Inspection of aggregated failure paradigms for PVTs only and SVTs only sheds light on whether overall performance validity (i.e., in accordance with Slick criteria) can be predicted with general equivalence by employing SVTs only compared to PVTs only or PVTs plus SVTs. In the aggregated SVT-only failure paradigm, criteria cutoffs of two or more, three or more, and four SVT failures yielded Sn equal to 0.319, 0.261, and 0.130, and Sp equal to 0.893, 0.939, and 0.990, respectively. Additionally, OC percentages were 74.3, 76.2, and 76.6, respectively. In contrast, the aggregated PVT-only failure paradigm exhibited better classification statistics overall. Specifically, cutoff criteria of PVT failures of two or more, three or more, four or more,

and five or more resulted in Sn of 0.783, 0.609, 0.493, and 0.319, and Sp of 0.842, 0.913, 0.970, and 0.990, with OC percentages of 82.6, 83.4, 84.5, and 81.5, respectively. Among the different PVT- and SVT-only failure paradigms, the highest OC percentage and, indeed, the false positive rate closest to 0.10 belonged to the cutoff of three or more PVT failures (Sn = 0.609, Sp = 0.913, OC% = 84.5). Comparatively, the best classification statistics of the combined PVT/SVT failure paradigm were achieved when applying cutoffs of four or more failures (Sn = 0.623, Sp = 0.883, OC% = 81.5) or five or more failures (Sn = 0.536, Sp = 0.959, OC% = 84.9). While it is debatable as to which PVT/SVT or PVT-only paradigm demonstrates the best overall ratio of classification statistics in the current study, it is evident that the SVT-only failure paradigm yielded the least desirable results.

Logistic regression was conducted using all 11 PVTs/SVTs as continuous independent variables to predict Valid/Invalid group membership. The purpose of this analysis was to serve as an internal validity check for the accuracy of the aggregated PVT/SVT failure paradigms in the study sample. Results from this logistic regression indicated a Sn equal to 0.667, a Sp equal to 0.944, and an OC percentage equal to 87.2. While this combination of classification statistics was slightly better than any individual paradigm of aggregated PVT/SVT failures, overall predictive accuracy was largely commensurate with criterion cutoffs of failure of four or more PVTs/SVTs (Sn = 0.623, Sp = 0.883, OC% = 81.5) and five or more PVT/SVT failures (Sn = 0.536, Sp = 0.959, OC% = 84.9).

A separate logistic regression was also performed with only the seven PVTs used as continuous independent variables, yielding Sn of 0.652, Sp of 0.944, and OC percentage of 86.8. Again, results were broadly comparable with PVT-only failure paradigm cutoffs of three or more failures (Sn = 0.609, Sp = 0.913, OC% = 83.4) and four or more failures (Sn = 0.493, Sp =

0.969, OC% = 84.5).

Predictive power statistics were also computed for individual PVT and SVT measures, as well as for aggregated failure paradigms. Positive predictive power (PPP) is defined as the probability of the presence of the disorder (i.e., Invalid Group membership in this case) given a positive test finding, and it takes into account base rate (BR) of the disorder. It is calculated by the following formula: (Base Rate x Sn)/{[(Base Rate x Sn)+[(1 – Base Rate) x (1-Sp)]}. As Greve and Bianchini (2004) noted the key to improving the diagnostic accuracy of malingering tests is to improve PPP, calculations of PPP were completed (Table 7) with various base rate values representing approximations of base rates of malingered neurocognitive dysfunction expected in litigation samples (BR ~ 0.4), non-clinical samples (BR ~ 0.1), and two values in between (0.2 and 0.3). Notably, the highest PPP for an individual measure in the predictive model across the different base rates was VRRec, which had PPP values of 0.867 for a BR of 0.4 and 0.522 for a BR of 0.1 (Table 7). In these same litigation (0.4) and non-clinical (0.1) estimated base rates, the aggregated failure paradigms of four or more and five or more PVTs/SVTs had PPP values equal to 0.780 (0.4) and 0.897 (0.4), and 0.371 (0.1) and 0.593 (0.1), respectively. A cutoff of three or more PVT failures had PPP values of 0.824 (0.4) and 0.438 (0.1) (Table 8)

Table 7

Likelihood Ratios and Positive Predictive Power Statistics for Individual PVTs/SVTs

Measures	Cutoff	LR+	LR-	PPP			
				BR = 0.4	BR = 0.3	BR = 0.2	BR = 0.1
Coding	≤37	4.26	0.10	0.740	0.646	0.516	0.321
Trails B	≥147	4.04	0.11	0.729	0.634	0.502	0.310
LMRec	≤19	4.00	0.13	0.727	0.632	0.500	0.308
VPARec	≤32	4.26	0.13	0.740	0.646	0.516	0.321
VRRec	≤3	9.81	0.06	0.867	0.808	0.710	0.522

Measures	Cutoff	LR+	LR-	PPP			
				BR = 0.4	BR = 0.3	BR = 0.2	BR = 0.1
BDAE-CIM	≤8	3.98	0.15	0.726	0.630	0.499	0.306
Visual Puzzles	≤7	5.30	0.08	0.779	0.694	0.570	0.371
Fr	≥109	3.51	0.09	0.701	0.601	0.467	0.281
FPr	≥81	3.22	0.08	0.682	0.580	0.446	0.263
FBSr	≥85	3.41	0.11	0.694	0.594	0.460	0.275
RBS	≥99	4.78	0.13	0.761	0.672	0.544	0.347

Table 8

Positive Predictive Power Values for Aggregated Failure Paradigms

Paradigm	Cutoff	PPP			
		BR = 0.4	BR = 0.3	BR = 0.2	BR = 0.1
PVTs/SVTs	≥3 failures	0.749	0.657	0.528	0.332
	≥4 failures	0.780	0.695	0.570	0.371
	≥5 failures	0.898	0.849	0.767	0.593
PVTs only	≥3 failures	0.824	0.750	0.637	0.438
	≥4 failures	0.915	0.873	0.801	0.641
	≥5 failures	0.954	0.931	0.887	0.776
SVTs only	≥2 failures	0.665	0.561	0.427	0.248
	≥3 failures	0.740	0.646	0.516	0.321
	= 4 failures	0.895	0.846	0.762	0.587

Also displayed in Table 7 are positive likelihood ratios (LR+) and negative likelihood ratios (LR-), as well as PPP, for all individual PVTs/SVTs. Notably, Grimes and Schultz (2005) outlined benchmarks for how LR+ values applying to pretest probabilities between 10 percent and 90 percent increase posttest probabilities. These benchmarks indicated LR+ values of two, five, and 10 increase posttest probability by 15, 30, and 45 percent, respectively. The highest LR+ of any individual PVT or SVT, by far, was equal to 9.81 for VRRec, suggesting a failure on this measure at the specified cutoff would indicated an approximate 45 percent increase in the posttest probability of belonging to the invalid group in this sample. Indeed, VRRec exhibited

the best predictive value of any individual measure across numerous diagnostic classification statistics.

CHAPTER 4

DISCUSSION

Results from the current study provided support for aggregating failures across different combinations of PVTs/SVTs (or PVTs only), with a per test false positive rate of 10 percent, as a method for predicting suboptimal effort in a monetarily incentivized forensic setting. Specifically, the aggregated failure paradigms with cutoffs of four or more PVT/SVT failures (Sn = 0.623, Sp = 0.883, OC% = 81.5) and five or more PVT/SVT failures (Sn = 0.536, Sp = 0.959, OC% = 84.9) yielded comparable predictive accuracy to the logistic regression using all 11 PVTs/SVTs as continuous independent variables (Sn = 0.667, Sp = 0.944, OC% = 87.2). Furthermore, results of a logistic regression with PVTs only (Sn = 0.652, Sp = 0.944, OC% = 86.8) were most similar to the PVT-only failure paradigms using cutoffs of three or more failures (Sn = 0.609, Sp = 0.913, OC% = 83.4) and four or more failures (Sn = 0.493, Sp = 0.969, OC% = 84.5). While failed SVTs generally correlated with failed PVTs in this sample, the power to predict malingering with SVTs alone was appreciably less than the use of solely embedded PVTs or a combination of PVTs and SVTs. Overall results showed that using multiple PVTs or PVTs/SVTs improved diagnostic classification accuracy over any one measure, arguably with the exception of VRRec, which exhibited singularly impressive predictive data and was clearly the most outstanding individual test out of all the embedded PVTs or SVTs in the present study.

Regarding Hypothesis (1), these findings were, in fact, aligned with expectations (based on results from Larrabee [2019]) that a combination of three or more failures at a per test specificity of 90 percent would best compare to the logistic regression classification accuracy. However, three or more failures was most comparable to the regression formula within the PVT-only paradigm, while four/five or more failures was nearest in accuracy to the regression within

the PVT/SVT paradigm, which was technically the tabulation model utilized in Larrabee (2019). The relative increase in the number of test failures required to most closely approximate the regression formula and echo findings from Larrabee (2019) is likely due to one main reason. Namely, the aggregated PVT/SVT failure paradigm used in the current study included four SVTs, while the model in Larrabee (2019) included only one SVT (i.e., the Meyers Index). Though the current study clearly showed that while the MMPI-2-RF response bias indices analyzed did offer some value in differentiating valid from invalid group members, it clearly performed worse than the embedded PVTs. Thus, the relatively larger percentage of SVTs in the current study's PVT/SVT failure paradigm (i.e., 4 out of 11, 36%) compared the paradigm in Larrabee (2019) (i.e., 1 out of 11, 9%) negatively impacted the former combination model by inflating the average number of failures needed to reach satisfactory prediction accuracy. Indeed, when the four SVTs were removed from the equation, the required cutoff of PVT failures to reach accuracy analogous to the logistic regression dropped to three or more failures.

Regarding Hypothesis (2), the BDAE-CIM and the three WMS-IV recognition subtests were among the measures with the most desirable diagnostic classification accuracies, as expected. Specifically, VRRec, VPARec, and BDAE-CIM stood out as the three best individual predictors, with Sn all close to or above 50 percent with Sp near 90% and the largest effect size differences between the valid and invalid groups (d >1). However, interestingly, with respect to the embedded WMS-IV measures, VRRec and VPARec notably demonstrated superior predictive statistics compared to LMRec, including: Sn (0.551 and 0.478 vs. 0.449), AUC (0.822 and 0.816 vs. 0.767), valid group skewness (-0.859 and -2.731 vs. -0.277; i.e., higher skewness in valid vs. invalid groups has been shown to be a common characteristic in well validated forced-choice validity measures), and effect size between valid/invalid groups (-1.26 and -1.08

vs. -0.97). Though LMRec was still largely adequate as an embedded PVT, these compelling differences warrant future investigations into the discrepancies between WMS-IV recognition measures as PVTs.

Regarding Hypothesis (3), predictions that the base rate of definite or probable MND in the study sample would mimic expectations from comparable groups in the in literature were not accurate. Specifically, current findings demonstrated a 26 percent base rate of definite or probable MND for the retired NFL players included in this study, though it was hypothesized to approximately align with the 38 percent figure widely accepted as the base rate of malingering in "high stakes" neuropsychological evaluations involving litigating mild TBI samples (Mittenberg, Patton, Canyock, and Condit, 2002). There are several points of consideration that might explain this notably lower ratio. First, to the author's knowledge, the current study represents the first known research establishing a base rate for MND in a sample of retired NFL players involved in the NFL BAP concussion settlement. Therefore, there was no true precedent from previous studies to which precise comparisons could be made. Second, it is possible that individuals in the sample were less incentivized by the external monetary awards possible in the settlement, as average career earnings for former NFL players are significantly higher than the national mean, regardless of era played. Third, the actuarial tables in the NFL BAP are dependent upon the number of concussions or concussive-like events, which, in turn, were estimated based on the number of games played, with a minimum of four games required for inclusion (i.e., a determination was made that if a retiree played in at least four games in a season, they were highly likely to have experienced at least one concussion-like event characterized by an episode of altered mental status). Thus, if a player was evaluated in the BAP, it means the player probably had incurred a concussion at some point in his NFL career, according to the settlement.

Accordingly, a reasonable argument could be made that many retired players in the sample were genuinely concerned about accurately assessing their brain health, given that they most likely had suffered at least one concussion, historically. This, consequently, might have reduced the rate of malingering compared to litigating TBI samples in which it can be inferred that not all individuals had necessarily incurred head injuries simply based off of their inclusion in a litigating TBI sample. Fourth, observed MND base rates might have been lower than expected due to the threat of coaching. This point is especially pertinent given the fixed battery of tests administered to thousands of players since 2014, which provides ample time for test security to have been compromised and evaluation coaching to have occurred. Last, the below expected base rates could also reflect limitations in the validity tests themselves (i.e., TOMM, ACS WC, MSVT, and RDS), as clinical neuropsychologists were wed to these specific measures and unable to employ other well validated PVTs (e.g., the Rey-15), which might have improved detection of feigned neurocognitive impairment.

Findings from the current study also successfully demonstrated specific cutoff scores for use of cognitive measures and the MMPI-2-RF as embedded validity measures in samples of retired NFL players participating in BAP evaluations. Further, since the PVTs used for classification prediction were standard clinical tools originally designed to measure certain neuropsychological constructs, the present results add valuable information to the literature for new cutoff score considerations for both previously research tests (e.g., BDAE-CIM, WMS-IV Recognition subtests, and Trails B) and for tests less studied for their potential to be utilized as embedded PVTs (e.g., WAIS-IV Coding and Visual Puzzles). While per test cutoff criteria were adjusted to be as close to the highly prioritized 10 percent false positive rate as possible, which applies specifically to the current sample, thereby limiting generalizability to other samples, it is

the hope of the author that these findings might serve as a foundation from which future studies can build upon for valid extrapolation to similar populations. While specific clinical histories for retired player within this archival dataset were inaccessible for the present study, comparable populations would include individuals with histories of chronic pain, psychiatric comorbidities, and repeated concussions who undergo neuropsychological evaluations in significantly incentivized forensic contexts.

In summary, the most salient findings from the current study provided evidence for using diagnostic paradigms built on aggregated failures of individual validity tests for assistance in confidently classifying definite or probable MND, in conjunction with other facets of Slick criteria (i.e., namely, the presence of a clear external incentive). This research is valuable for research and clinical purposes. For the latter, this diagnostic approach is helpful to clinicians who more commonly employ flexible versus fixed batteries without a set list of measures, and thus, do not always administer the same combinations of tests. Despite the diversity of assessment battery approaches that exist, some of the most popular measures used in neuropsychological evaluations were analyzed in this study, which maximizes the applicability of the current findings. Moreover, this study emphasized variety in the types of tests selected in order to echo the comprehensive battery approach often used in forensic evaluations. This was achieved by including a wide range of neuropsychological constructs for which selected tests were designed to measure, as well as timed and untimed test structures, forced choice and non-forced choice formats, and embedded symptom validity indices on a self-report measure.

This study was strongly influenced by the work of Larrabee (2019), findings from which the current research sought to expand upon and limitations of which targets for improvement. These goals were achieved by replicating broadly similar recommendations for using at least

three aggregated PVT failures at a per test specificity of 90 percent as a cutoff criterion and with respect to using multiple PVTs as sources for initial classification of MND and non-MND groups, which adhered to the "gold standard" diagnostic scheme proposed by Slick et al. (1999). However, the present study still has several limitations, some of which were also acknowledged by Larrabee (2019). First, cutoff scores for embedded PVTs/SVTs were specifically adjusted to reach false positive rates as close to 0.10 as possible for the sample, but validity measures used for original group delineation (i.e., TOMM, ACS WC, MSVT, and RDS) were not adjusted in parallel. This restricts the applicability of the particular cutoff scores to populations beyond the current sample and limits clinical extrapolation, as clinicians do not typically have large explicit samples akin to every client from which cutoff scores can be adjusted accordingly. Nonetheless, the principle of utilizing recommended cutoffs at 90 percent specificity in aggregate as a diagnostic paradigm for MND classification was highlighted by the current findings and can still be used as a guide when following published guidelines for cutoff criteria in specific clinical populations. Another limitation is the lack of clinical context for the retired players in the sample (i.e., with the exception of a probable history of concussions) as a consequence of the restrictions inherent in this archival dataset. Yet another limitation was the exclusion of retirees over the age of 69. While their exclusion helped mitigate the confounding effects of higher rates of neurodegenerative disorders in the sample, which would have likely increased the number of false positives (a high priority for effort test validation research), it also reduces the scope of investigation into the full BAP population, subsequently diminishing valuable insight. Finally, while the inclusion of Coding and Visual Puzzles as tests in the prediction paradigm provided some new perspective into their potential as embedded PVTs, as research in this area is narrow, their selection into the model was limited in theoretical justification and might have

compromised model parsimony, ultimately decreasing diagnostic predictive power.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

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

Ardolf, B. R., Denney, R. L., & Houston, C. M. (2007). Base rates of negative response bias and malingered neurocognitive dysfunction among criminal defendants referred for neuropsychological evaluation. *The Clinical Neuropsychologist*, *21*(6), 899-916. https://doi.org/10.1080/13825580600966391

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*(1), 145-159.

Bauer, L., O'Bryant, S. E., Lynch, J. K., McCaffrey, R. J., & Fisher, J. M. (2007). Examining the test of memory malingering trial 1 and word memory test immediate recognition as screening tools for insufficient effort. *Assessment*, *14*(3), 215-222.

Berry, D. T., & Nelson, N. W. (2010). DSM-5 and malingering: A modest proposal. *Psychological Injury and Law*, *3*(4), 295-303. 10.1007/s12207-010-9087-7

Berry, D. T., & Schipper, L. J. (2007). Detection of feigned psychiatric symptoms during forensic neuropsychological examinations. *Assessment of malingered neuropsychological deficits*, 226-263.

Bianchini, K. J., Curtis, K. L., & Greve, K. W. (2006). Compensation and malingering in traumatic brain injury: a dose-response relationship?. *The Clinical Neuropsychologist*, *20*(4), 831-847. https://doi.org/10.1080/13854040600875203

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*(4), 632-640. doi:10.1017/S1355617712000252

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain injury*, *28*(13-14), 1623-1638. https://doi.org/10.3109/02699052.2014.947627

Blackstone, W. (1844). Commentaries on the Laws of England: Book I: Of the Rights of Persons. Jazzybee Verlag.

Boake, C., McCauley, S. R., Levin, H. S., Contant, C. F., Song, J. X., Brown, S. A., ... & Merritt, S. G. (2004). Limited agreement between criteria-based diagnoses of postconcussional syndrome. *The Journal of neuropsychiatry and clinical neurosciences*, *16*(4), 493-499. https://doi.org/10.1176/jnp.16.4.493

Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P., & Singh, H. (2009). Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer disease and associated disorders*, *23*(4), 306. https://dx.doi.org/10.1097%2FWAD.0b013e3181a6bebc

Brennan, A. M., Meyer, S., David, E., Pella, R., Hill, B. D., & Gouvier, W. D. (2009). The vulnerability to coaching across measures of effort. *The Clinical Neuropsychologist*, *23*(2), 314-328. https://doi.org/10.1080/13854040802054151

Breting, L. G., & Sweet, J. (2013). Freestanding cognitive symptom validity tests: Use and selection in mild traumatic brain injury. In *Mild traumatic brain injury, symptom validity assessment and malingering* (pp. 145-158). Springer, New York.

Carone, D. A. (2008). Children with moderate/severe brain damage/dysfunction outperform adults with mild-to-no brain damage on the Medical Symptom Validity Test. *Brain Injury*, *22*(12), 960-971. https://doi.org/10.1080/02699050802491297

Carone, D. A., Bush, S. S., & Iverson, G. L. (2013). Providing feedback on symptom validity, mental health, and treatment in mild traumatic brain injury. *Mild traumatic brain injury; Symptom validity assessment and malingering*, 101-118.

Chen, Y. C., Smith, D. H., & Meaney, D. F. (2009). In-vitro approaches for studying blast-induced traumatic brain injury. *Journal of neurotrauma*, *26*(6), 861-876. https://dx.doi.org/10.1089%2Fneu.2008.0645

Davis, J. J., & Millis, S. R. (2014). Examination of performance validity test failure in relation to number of tests administered. *The Clinical Neuropsychologist*, *28*(2), 199-214.

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*(1), 133-152.

Erdodi, L. A., Tyson, B. T., Abeare, C. A., Lichtenstein, J. D., Pelletier, C. L., Rai, J. K., & Roth, R. M. (2016). The BDAE Complex Ideational Material—A measure of receptive language or performance validity?. *Psychological Injury and Law*, *9*(2), 112-120.

Essig, S. M., Mittenberg, W., Petersen, R. S., Strauman, S., & Cooper, J. T. (2001). Practices in forensic neuropsychology: Perspectives of neuropsychologists and trial attorneys. *Archives of Clinical Neuropsychology*, *16*(3), 271-291. https://doi.org/10.1016/S0887-6177(99)00065-7

Etherton, J. L., Bianchini, K. J., Greve, K. W., & Heinly, M. T. (2005). Sensitivity and specificity of reliable digit span in malingered pain-related disability. *Assessment*, *12*(2), 130-136. https://doi.org/10.1177/1073191105274859

Evans, R. W. (2010). Persistent Post-Traumatic Headache, Postconcussion Syndrome, and Whiplash Injuries: The Evidence for a Non-Traumatic Basis With an Historical Review. *Headache: The Journal of Head and Face Pain*, *50*(4), 716-724. https://doi.org/10.1111/j.1526-4610.2010.01645.x

Fann, J. R., Katon, W. J., Uomoto, J. M., & Esselman, P. C. (1995). Psychiatric disorders and functional disability in outpatients with traumatic brain injuries. *The American journal of psychiatry, 152*(10), 1493. https://search.proquest.com/docview/220457337?pq-origsite=gscholar

Faul, M., & Coronado, V. (2015). Epidemiology of traumatic brain injury. In *Handbook of clinical neurology* (Vol. 127, pp. 3-13). Elsevier. https://doi.org/10.1016/B978-0-444-52892-6.00001-5

Frederick, R. I., & Speed, F. M. (2007). On the interpretation of below-chance responding in forced-choice tests. *Assessment, 14*(1), 3-11. https://doi.org/10.1177/1073191106292009

Gavett, B. E., Stern, R. A., & McKee, A. C. (2011). Chronic traumatic encephalopathy: a potential late effect of sport-related concussive and subconcussive head trauma. *Clinics in sports medicine*, *30*(1), 179-188. https://doi.org/10.1016/j.csm.2010.09.007

Gennarelli, T. A., & Graham, D. I. (2005). Neuropathology. *Textbook of traumatic brain injury*, 27-50.

Green, P. (2004). *Medical Symptom Validity Test (MSVT) for microsoft windows: User's manual*. Paul Green Pub..

Green, P. (2005). *Green's Word Memory Test for Microsoft Windows: User's manual*. Green's Publications Incorporated.

Green, P., Iverson, G. L., & Allen, L. (1999). Detecting malingering in head injury litigation with the Word Memory Test. *Brain Injury*, *13*(10), 813-819. https://doi.org/10.1080/026990599121205

Green, P., Rohling, M. L., Lees-Haley, P. R., & III, L. M. A. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain injury*, *15*(12), 1045-1060. https://doi.org/10.1080/02699050110088254

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, *6*(3), 218. http://psycnet.apa.org/doi/10.1037/1040-3590.6.3.218

Greve, K. W., & Bianchini, K. J. (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., Bianchini, K. J., & Doane, B. M. (2006). Classification accuracy of the Test of Memory Malingering in traumatic brain injury: Results of a known-groups analysis. *Journal of Clinical and Experimental Neuropsychology*, *28*(7), 1176-1190. https://doi.org/10.1080/13803390500263550

Greve, K. W., Binder, L. M., & Bianchini, K. J. (2009). Rates of below-chance performance in forced-choice symptom validity tests. *The Clinical Neuropsychologist*, *23*(3), 534-544.

https://doi.org/10.1080/13854040802232690

Greve, K. W., Etherton, J. L., Ord, J., Bianchini, K. J., & Curtis, K. L. (2009). Detecting malingered pain-related disability: Classification accuracy of the Test of Memory Malingering. *The Clinical Neuropsychologist*, *23*(7), 1250-1271.

Greve, K. W., Springer, S., Bianchini, K. J., Black, F. W., Heinly, M. T., Love, J. M., ... & 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.

Grimes, D. A., & Schulz, K. F. (2005). Refining clinical diagnosis with likelihood ratios. *The Lancet*, *365*(9469), 1500-1505.

Gualtieri, T., & Cox, D. R. (1991). The delayed neurobehavioural sequelae of traumatic brain injury. *Brain injury*, *5*(3), 219-232. https://doi.org/10.3109/02699059109008093

Haber, A. H., & Fichtenberg, N. L. (2006). Replication of the Test of Memory Malingering (TOMM) in a traumatic brain injury and head trauma sample. *The Clinical Neuropsychologist*, *20*(3), 524-532. https://doi.org/10.1080/13854040590967595

Hartman, D. E. (2009). Wechsler Adult Intelligence Scale IV (WAIS IV): return of the gold standard. *Applied neuropsychology*, *16*(1), 85-87.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference Participants 1. (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. http://dx.doi.org/10.1080/13854040903155063

Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia*, *42*(9), 1212-1222.

Holdnack, J. A., & Drozdick, L. W. (2009). Advanced clinical solutions for WAIS-IV and WMS-IV: Clinical and interpretive manual. *Texas: Pearson*.

Holdnack, J. A., Schoenberg, M. R., Lange, R. T., & Iverson, G. L. (2013). Predicting Premorbid Ability for WAIS–IV, WMS–IV and WASI–II. In *WAIS-IV, WMS-IV, and ACS* (pp. 217-278). Academic Press.

Iverson, G. L., Lange, R. T., Brooks, B. L., & Lynn Ashton Rennison, V. (2010). "Good old days" bias following mild traumatic brain injury. *The Clinical Neuropsychologist*, *24*(1), 17-37. https://doi.org/10.1080/13854040903190797

Kanazawa, I. (2001). How do neurons die in neurodegenerative diseases?. *Trends in molecular medicine*, *7*(8), 339-344. https://doi.org/10.1016/S1471-4914(01)02017-2

Kaufmann, P. M. (2009). Protecting raw data and psychological tests from wrongful disclosure: A primer on the law and other persuasive strategies. *The Clinical Neuropsychologist*,

23(7), 1130-1159. https://doi.org/10.1080/13854040903107809

Kohn, S. E., & Goodglass, H. (1985). Picture-naming in aphasia. *Brain and language*, *24*(2), 266-283.

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist*, *17*(3), 410-425. http://psycnet.apa.org/doi/10.1076/clin.17.3.410.18089

Larrabee, G. J. (Ed.). (2007). *Assessment of malingered neuropsychological deficits*. Oxford University Press.

Larrabee, G. J., Rohling, M. L., & Meyers, J. E. (2019). Use of multiple performance and symptom validity measures: Determining the optimal per test cutoff for determination of invalidity, analysis of skew, and inter-test correlations in valid and invalid performance groups. *The Clinical Neuropsychologist*, *33*(8), 1354-1372.

Legg, M. (2015). National football league players' concussion injury class action settlement. *Australian and New Zealand Sports Law Journal*, *10*(1), 47.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). Neuropsychological assessment (5th ed.). New York, NY, US: Oxford University Press.

Lindstrom Jr, W. A., Lindstrom, J. H., Coleman, C., Nelson, J., & Gregg, N. (2009). The diagnostic accuracy of symptom validity tests when used with postsecondary students with learning disabilities: A preliminary investigation. *Archives of Clinical Neuropsychology*, *24*(7), 659-669.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315-1316.

Marshall, P., Schroeder, R., O'Brien, J., Fischer, R., Ries, A., Blesi, B., & Barker, J. (2010). Effectiveness of symptom validity measures in identifying cognitive and behavioral symptom exaggeration in adult attention deficit hyperactivity disorder. *The Clinical Neuropsychologist*, *24*(7), 1204-1237. https://doi.org/10.1080/13854046.2010.514290

Mathias, C. W., Greve, K. W., Bianchini, K. J., Houston, R. J., & Crouch, J. A. (2002). Detecting malingered neurocognitive dysfunction using the reliable digit span in traumatic brain injury. *Assessment*, *9*(3), 301-308. https://doi.org/10.1177/1073191102009003009

McCrory, P., Meeuwisse, W., Johnston, K., Dvorak, J., Aubry, M., Molloy, M., & Cantu, R. (2009). Consensus statement on concussion in sport–the 3rd International Conference on concussion in sport, held in Zurich, November 2008. *Journal of Clinical Neuroscience*, *16*(6), 755-763. https://doi.org/10.1016/j.jocn.2009.02.002

McGrath, B. (2011). Does football have a future? The NFL and the concussion crisis. *New Yorker (New York, NY: 1925)*, 40.

Mez, J., Daneshvar, D. H., Kiernan, P. T., Abdolmohammadi, B., Alvarez, V. E., Huber, B. R., ... & Cormier, K. A. (2017). Clinicopathological evaluation of chronic traumatic encephalopathy in players of American football. *Jama*, *318*(4), 360-370. http://dx.doi.org/10.1001/jama.2017.8334

Millis, S. R., & Volinsky, C. T. (2001). Assessment of response bias in mild head injury: Beyond malingering tests. *Journal of Clinical and Experimental Neuropsychology*, *23*(6), 809-828.

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exeggeration. *Journal of clinical and experimental neuropsychology*, *24*(8), 1094-1102. https://doi.org/10.1076/jcen.24.8.1094.8379

Mittenberg, W., Theroux-Fichera, S., Zielinski, R., & Heilbronner, R. L. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale—Revised. *Professional Psychology: Research and Practice*, *26*(5), 491. doi:10.1037/0735-7028.26.5.491

Morgan, J. E., & Ricker, J. H. (Eds.). (2016). *Textbook of clinical neuropsychology*. Taylor & Francis.

Morris, J. C. (1997). Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International psychogeriatrics*, *9*(S1), 173-176. https://doi.org/10.1017/S1041610297004870

Nelson, H. E., & Willison, J. (1991). *National adult reading test (NART)*. Windsor: Nfer-Nelson.NFL concussion settlement: posted settlement program. (2018). Retrieved from https://www.nflconcussionsettlement.com/Docs/Posted_Settlement_FAQs.pdf

Nies, K. J., & Sweet, J. J. (1994). Neuropsychological assessment and malingering: A critical review of past and present strategies. *Archives of Clinical Neuropsychology*, *9*(6), 501-552. https://doi.org/10.1093/arclin/9.6.501

O'Bryant, S. E., Lacritz, L. H., Hall, J., Waring, S. C., Chan, W., Khodr, Z. G., ... & Cullum, C. M. (2010). Validation of the new interpretive guidelines for the clinical dementia rating scale sum of boxes score in the national Alzheimer's coordinating center database. *Archives of Neurology*, *67*(6), 746-749. doi:10.1001/archneurol.2010.115

Omalu, B. I., Hamilton, R. L., Kamboh, M. I., DeKosky, S. T., & Bailes, J. (2010). Chronic traumatic encephalopathy (CTE) in a National Football League Player: Case report and emerging medicolegal practice questions. *Journal of forensic nursing*, *6*(1), 40-46. https://doi.org/10.1111/j.1939-3938.2009.01064.x

Oremus, M., Perrault, A., Demers, L., & Wolfson, C. (2000). Review of outcome measurement instruments in Alzheimer's disease drug trials: psychometric properties of global scales. *Journal of geriatric psychiatry and neurology*, *13*(4), 197-205. https://doi.org/10.1177/089198870001300404

Pankratz, L., & Erickson, R. C. (1990). Two views of malingering. *The Clinical*

Neuropsychologist, *4*(4), 379-389. https://doi.org/10.1080/13854049008401832

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology*, *31*(3), 206-230.

Randolph, C., Lansing, A. E., Ivnik, R. J., Cullum, C. M., & Hermann, B. P. (1999). Determinants of confrontation naming performance. *Archives of clinical neuropsychology*, *14*(6), 489-496.

Rogers, R. (1997). Current status of clinical methods. *Clinical assessment of malingering and deception*, *2*, 373-379.

Rogers, R. (2008). Detection strategies for malingering and defensiveness. *Clinical assessment of malingering and deception*, *3*, 14-35.

Rogers, R., Bender, S. D., & Johnson, S. F. (2011). A critical analysis of the MND criteria for feigned cognitive impairment: Implications for forensic practice and research. *Psychological Injury and Law*, *4*(2), 147-156. http://dx.doi.org/10.1007%2Fs12207-011-9107-2

Rogers, R., & Correa, A. A. (2008). Determinations of malingering: Evolution from case-based methods to detection strategies. *Psychiatry, Psychology and Law*, *15*(2), 213-223. https://doi.org/10.1080/13218710802014501

Rogers, R., Harrell, E. H., & Liff, C. D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review*, *13*(3), 255-274. https://doi.org/10.1016/0272-7358(93)90023-F

Rosenfeld, B., Sands, S. A., & Van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology*, *15*(4), 349-359. https://doi.org/10.1016/S0887-6177(99)00025-6

Ruiz, M. A., Drake, E. B., Glass, A., Marcotte, D., & van Gorp, W. G. (2002). Trying to beat the system: Misuse of the Internet to assist in avoiding the detection of psychological symptom dissimulation. *Professional Psychology: Research and Practice*, *33*(3), 294. http://psycnet.apa.org/doi/10.1037/0735-7028.33.3.294

Schroeder, R. W., Martin, P. K., & Odland, A. P. (2016). Expert beliefs and practices regarding neuropsychological validity testing. *The Clinical Neuropsychologist*, *30*(4), 515-535.

Sharland, M. J., & Gfeller, J. D. (2007). A survey of neuropsychologists' beliefs and practices with respect to the assessment of effort. *Archives of Clinical Neuropsychology*, *22*(2), 213-223. https://doi.org/10.1016/j.acn.2006.12.004

Slick, D. J., & Sherman, E. M. (2012). Differential diagnosis of malingering and related clinical presentations. *Pediatric forensic neuropsychology*, 113-135.

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, *13*(4), 545-561. https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT545

Slick, D. J., Tan, J. E., Sherman, E. M., & Strauss, E. (2010). 39 Malingering and Related Conditions in Pediatric Populations. *Handbook of pediatric neuropsychology*, 457.

Sweet, J. J., Goldman, D. J., Breting, G., & Leslie, M. (2013). Traumatic brain injury: guidance in a forensic context from outcome, dose–response, and response bias research. *Behavioral Sciences & the law*, *31*(6), 756-778. https://doi.org/10.1002/bsl.2088

Sweet, J. J., Meyer, D. G., Nelson, N. W., & Moberg, P. J. (2011). The TCN/AACN 2010 "salary survey": Professional practices, beliefs, and incomes of US neuropsychologists. *The Clinical Neuropsychologist*, *25*(1), 12-61. https://doi.org/10.1080/13854046.2010.544165

Tombaugh, T. N. (1996). Test of memory malingering: TOMM. New York/Toronto: MHS.

Victor, T. L., Boone, K. B., Serpa, J. G., Buehler, J., & Ziegler, E. A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist*, *23*(2), 297-313.

Vitacco, M. J., Jackson, R. L., Rogers, R., Neumann, C. S., Miller, H. A., & Gabel, J. (2008). Detection strategies for malingering with the Miller Forensic Assessment of Symptoms Test: A confirmatory factor analysis of its underlying dimensions. *Assessment*, *15*(1), 97-103. doi:10.1.1.993.1761

Walters, G. D., Rogers, R., Berry, D. T., Miller, H. A., Duncan, S. A., McCusker, P. J., ... & Granacher Jr, R. P. (2008). Malingering as a categorical or dimensional construct: The latent structure of feigned psychopathology as measured by the SIRS and MMPI-2. *Psychological Assessment*, *20*(3), 238. 10.1037/1040-3590.20.3.238

Wood, R. L., & Rutterford, N. A. (2006). The effect of litigation on long term cognitive and psychosocial outcome after severe brain injury. *Archives of clinical neuropsychology*, *21*(3), 239-246. https://doi.org/10.1016/j.acn.2005.12.004

Youngjohn, J. R., Lees-Haley, P. R., & Binder, L. M. (1999). Comment: Warning malingerers produces more sophisticated malingering. *Archives of Clinical Neuropsychology*, *14*(6), 511-515. https://doi.org/10.1016/S0887-6177(98)00049-3

Zaloshnja, E., Miller, T., Langlois, J. A., & Selassie, A. W. (2008). Prevalence of long-term disability from traumatic brain injury in the civilian population of the United States, 2005. *The Journal of head trauma rehabilitation*, *23*(6), 394-400. 10.1097/01.HTR.0000341435.52004.ac