1-1-2010

Assessment Of Memory Function And Effort Using The Wechsler Memory Scale - 4th Edition

Justin B. Miller
Wayne State University,

ASSESSMENT OF MEMORY FUNCTION AND EFFORT USING THE WECHSLER MEMORY SCALE – 4TH EDITION

by

JUSTIN MILLER, M.A.

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2011

MAJOR: PSYCHOLOGY (Clinical)

Approved by:

____

Advisor Date

____

____

____

____

DEDICATION

This work is dedicated to family.

Without you, the following pages would be blank.

ACKNOWLEDGMENTS

Despite only a single name on the title page, completion of this project would not have been possible without the efforts of many individuals – far too many to name – who have been instrumental in the research process from beginning to end. Although there can be no way to name everyone, several people played an integral role. First and foremost, the guidance and support from my advisors, Dr. Lisa J. Rapport and Dr. R. Douglas Whitman, and dissertation committee members, Dr. Robin A, Hanks, and Dr. Bradley N. Axelrod, has been crucial to the development, execution, and completion of a project worthy of the dissertation status. In addition, the efforts of Carole Koviak, R.N. during recruitment and Jesse R. Bashem during data collection, dramatically expedited assembly of a completed database. Special thanks are also extended to Dr. Scott R. Millis for his generous help during analysis. The Del Harder Foundation is also acknowledged for its generosity in funding student research; without such funding, the present project would not have gotten off the ground. In closing, to all those previously mentioned, and to all those who were not (but certainly not forgotten), I owe my deepest gratitude.

TABLE OF CONTENTS

CHAPTER 1

BACKGROUND

Traumatic brain injury (TBI) is a major health problem and one of the leading causes of disability claims in those under 40 with approximately 1.4 million new injuries each year (Draper & Ponsford, 2008; Langlois, Rutland-Brown, & Thomas, 2006). Approximately 80,000 – 90,000 TBI survivors demonstrate lasting impairments in a multitude of domains (Langlois et al., 2006). Memory impairment is common after TBI (Dikmen, Machamer, Winn, & Temkin, 1995; Fleming, Riley, Gill, Gullo, Strong, & Shum, 2008; Gronwall & Wrightson, 1981; McAllister, 1992; Tabaddor, Mattis, & Zazula, 1984; Vakil, 2005) and may last years after injury (Draper & Ponsford, 2008; Tabaddor et al., 1984). Understanding the nature and severity of these impairments is paramount for TBI survivors, clinicians, and researchers. Cognitive evaluations post TBI typically include a formal assessment of memory using well developed standardized measures. Unfortunately, even the most psychometrically sound measures are vulnerable to variability that is influenced by the level of effort put forth by the examinee. Less-than-full effort may be put forth by examinees for many reasons, but the common end result is that the obtained results are of questionable validity and accuracy. Although clinicians could simply ask examinees about the level of effort exerted during testing, their responses may be inaccurate due to conscious or unconscious reasons, and if an examinee is intentionally faking, it is highly unlikely they would willingly divulge such information. As a result, a number of stand-alone measures and empirically-derived prediction models have been developed to detect suboptimal effort. Historically, detection of suboptimal effort also has employed prediction models using embedded measures derived from tests commonly used in a comprehensive cognitive battery, such as the Wechsler Memory Scale (e.g., Ord, Greve, & Bianchini, 2008). The aims of the present study were to (1) investigate the ability of the newest revision of the Wechsler Memory Scale (4th Edition) to distinguish poor performance due to suboptimal effort among TBI dissimulators from poor performance due to actual TBI, and (2) develop an empirically-derived prediction formula for use in clinical practice.

Section 1.1 - Assessment of Memory

Memory functioning generally refers to the ability to retain learned information over time and is one of several primary domains of cognitive ability typically evaluated in the context of a comprehensive neuropsychological assessment (Lezak, Howieson, Loring, Hannay, & Fischer, 2004; Pereira, 2007). It is important to note that memory is typically not viewed as a unitary construct but rather as a cluster of related abilities based on the duration of time that information is retained (e.g., short-term vs. long-term) and the content that is stored (e.g., long-term verbal memory, etc.; Baddeley, 2002; Pereira, 2007; Vakil, 2005). Furthermore, memory is also composed of multiple processes that are typically linear in progression, any of which can be disrupted to impair memory functioning. The most commonly referenced processes include encoding, storage and retrieval (Baddeley, 1997; Pereira, 2007), and each of these must be assessed in order to provide a comprehensive understanding of a clients memory functioning.

The Wechsler Memory Scale (WMS) is a battery of subtests frequently administered in a neuropsychological assessment (Rabin, Barr, & Burton, 2005) and is designed to evaluate multiple aspects of learning and memory. The measure originally appeared for clinical use in 1945 and was first published as an article in the Journal of Psychology (Wechsler, 1945). It has recently been revised and updated to the current fourth edition (WMS-IV), and it was substantially revised from the previous edition. Administration of this new edition in its entirety generates primary index scores for immediate and delayed memory as well as secondary indexes of visual working memory, auditory memory, and visual memory (Pearson Education, 2008). In the 4th edition, only three out of seven subtests were retained from the previous version whereas the remaining four tests are new.

Section 1.2 – Memory Functioning in Traumatic Brain Injury

Following traumatic brain injury, cognitive deficits – particularly in relation to learning and memory – have been shown using a variety of assessment techniques including caregiver report (e.g., McKinlay, Brooks, Bond, Martinage, & Marshall, 1981), self-report (e.g., Boake, Freeland, Ringholz, Nance, & Edwards, 1995; Cicerone & Kalmar, 1995), and formal neuropsychological assessment (e.g., Dikmen et al., 1995; Draper & Ponsford, 2008). At 1 year post injury, some of the most commonly cited difficulties include slowed thinking and reduced processing speed, impaired attention, and memory deficits in addition to general cognitive impairment, all of which demonstrated a relationship with injury severity as measured by time to follow command (Dikmen et al., 1995). Considering that memory is a higher-order cognitive functioning in that it requires lower-order cognitive abilities such as attention and processing speed – skills that are commonly affected by TBI – it is no surprise that memory impairment is frequently observed among TBI survivors.

In a longitudinal survey of TBI survivor caregivers, poor memory was one of the most commonly cited deficits (McKinlay et al., 1981). Using the original version of the WMS, Brooks (1976) demonstrated that in comparison to a neurologically intact control sample, individuals with history of TBI not only demonstrated severe memory deficits, but also that the severity of their deficits were predicted by the length of post-traumatic amnesia, a finding which closely paralleled that of Dikmen et al. (1999) nearly two decades later. Subsequent research demonstrated that the rate of memory recovery following traumatic brain injury was predicted by length of post-traumatic disorientation (Gronwall & Wrightson, 1981; Parker & Serrats, 1976). More recently, several studies have emerged demonstrating the lasting cognitive and functional impairments associated with TBI, even as long as 30 years post injury (Himanen, Portin, Isoniemi, Helenius, Kurki, & Tenovuo, 2006; Hoofien, Gilboa, Vakil, & Donovick, 2001). In a recent longitudinal study focusing on cognitive ability at 10-year follow up, processing speed, memory, and executive functioning were significantly lower in a TBI group in comparison to demographically matched controls (Draper & Ponsford, 2008). It is important to note that the majority of cases studied by Draper and Ponsford were moderate to severe injuries as lasting impairments are not typical of mild traumatic brain injury.

Approximately 70% of all traumatic brain injuries are classified as “mild”, which typically is defined as a Glasgow Coma Scale (GCS) score of 13 or greater and a time-to-follow commands of 1 hour or less (Dikmen et al., 1995; Larrabee, 2005). In nearly all cases of uncomplicated mild injury, symptoms have completely resolved by approximately 3 months post injury (Binder, 1986; Binder & Rohling, 1996; Binder, Rohling, & Larrabee, 1997; Dikmen et al., 1995; Larrabee, 2005). Although some individuals do demonstrate persisting deficits, particularly in cases of complicated mild injury (Kashluba, Hanks, Casey, & Millis, 2008; Larrabee, 2005), the greater majority of mild TBI cases do not. Interestingly, however, mild TBI represents the most frequently referred case in forensic neuropsychology (Larrabee, 2005; Ruff & Richardson, 1999). These evaluations commonly attempt to determine whether any impairments resulting from a neurological insult are present (Binder, 1997; Sherer & Madison, 2005), and when one considers the prevalence of persisting deficits, the validity of a substantial number of injury claims is brought into question.

Unlike mild injuries, persisting cognitive impairments, including memory deficits, are present in a majority of cases for moderate and severe TBI survivors. The acute recovery phases are typically characterized by periods of confusion, attentional difficulties, and memory dysfunction. This cluster of symptoms closely paralleling a delusional state is frequently referred to as posttraumatic confusion or amnesia (PTA; Ahmed, Bierley, Sheikh, & Date, 2000; Forrester, Encel, & Geffen, 1994; Larrabee, 2005; Sherer & Madison, 2005) and is a good predictor of functional outcome following injury (Boake, Millis, High, Delmonico, Kreutzer, Rosenthal, Sherer, & Ivanhoe, 2001; Brown, Malec, McClelland, Diehl, Englander, & Cifu, 2005; De Guise, Leblanc, Feyz, & Lamoureux, 2005; Nakase-Richardson, Yablon, & Sherer, 2007). Length of PTA can also aid in injury classification (Shores, Lammel, Hullick, Sheedy, Flynn, Levick, & Batchelor, 2008) when used in conjunction with other indicators (e.g., intracranial abnormalities, positive neuroimaging).

Frequently following moderate and severe injuries, memory dysfunction is typified by both retrograde and anterograde amnesia in the acute recovery phases (Levin, 1992; Sherer & Madison, 2005). In the post-acute phase of recovery, specific patterns of memory functioning have been identified (Millis & Ricker, 1994) as well as memory impairment subtypes (Curtiss, Vanderploeg, Spencer, & Salazar, 2001), but a careful review of the literature suggests that memory functioning following moderate to severe TBI is highly variable. For example, although some have cited deficits in the encoding of new memories (Roche, Moody, Szabo, Fleming, & Shum, 2007), others have suggested that the underlying memory dysfunction observed in cases of TBI is more likely a consolidation deficit (Vanderploeg, Crowell, & Curtiss, 2001). More specifically, Vanderploeg et al. (2001) observed comparable learning rates between TBI survivors and normal controls but the TBI group demonstrated a much more rapid rate of forgetting, suggesting consolidation deficiencies. Furthermore, both groups benefited equally from retrieval cues implying that retrieval abilities were comparable between groups.

Curtiss, Vanderploeg, Spencer, and Salazar (2001) identified three distinct clusters of memory deficit deficits including impairments in retrieval, consolidation, and retention while also finding a subgroup of TBI survivors that demonstrated intact memory performance. Interestingly though, those demonstrating relatively preserved verbal learning abilities also appeared to adopt a more passive encoding strategy (e.g., reliance on serial position vs. semantic clustering) similar to the findings of Millis and Ricker (1994). Curtiss et al. (2001) also found evidence that even in their most impaired sample, immediate memory as indicated by digit span performance was within normal expectations. Although injury severity was not a significant predictor of memory deficit in their sample, they were also careful to point out that location of injury may have an important role in determining specific memory deficits. It is suspected that this finding is a significant factor contributing to the variability of post-TBI memory profiles within the literature.

Section 1.3 – Malingering

A fundamental assumption of clinical assessment is that clients and practitioners share similar goals in the assessment process (Rogers, 1997a), and moreover that the assessment is taken seriously by both parties. Unfortunately, however, this is not always the case. With the increasing prevalence of forensic evaluations and the potential for financial remuneration following demonstration of cognitive impairment, the temptation to feign dysfunction on formal testing has also risen. In addition to the quality of measures chosen, and the proficiency of the psychologist, the validity of results obtained in an evaluation of memory functioning are also highly dependent upon the level of effort exerted during the evaluation by the examinee. Several terms have been developed to describe situations where an examinee has greatly exaggerated their symptoms or intentionally put forth suboptimal effort during an evaluation (Larrabee, 2007) including “suspect” effort (Babikian, Boone, Lu, & Arnold, 2006) and “incomplete” effort (Axelrod, Fichtenberg, Millis, & Wertheimer, 2006). The phenomenon that they each refer to, however, is highly similar; a subset of behavior encompassed under this rubric includes the psychiatric disorder “malingering”.

According to the most recent version of the Diagnostic and Statistical Manual of Mental Disorders (Association, 2000), malingering is “the *intentional* production of false or grossly exaggerated symptoms that are motivated by external incentives". It is important to point out that effort is not a simple bifurcated phenomenon with only good and poor effort (Iverson, 2006) and also that malingering is distinct from other conditions where symptom exaggeration is a prominent factor such as factitious disorder and conversion disorder. Conversion disorder is the most discrepant from malingering in that the incentive is internal (e.g., psychological), and the exaggeration of symptoms is not under voluntary control. In factitious disorder, the fabrication of symptomatology is a conscious, voluntary effort, but there is no direct external incentive such as that which is found in malingering. Although closely related, malingering is also different from simple provision of poor effort without the intent to obtain some form of incentive (Iverson, 2006). The distinction between intentional exaggeration and malingering is clearly a very fine line, as is the distinction between poor effort and intentional feigning.

Although technically listed as a V-Code in the DSM-IV-TR (V65.2), several unofficial sets of classification schemes have been developed by practicing clinicians and researchers in order to facilitate communication, classification and research of malingering (e.g., Greiffenstein, Gola, & Baker, 1995; Rogers, 1990). Developed specifically for use in neuropsychology, one of the most commonly employed are the diagnostic criteria for malingered neurocognitive deficit (MND) proposed by Slick, Sherman and Iverson (1999). Diagnosis of MND is a multifaceted assessment method taking into consideration behavioral self-report, observations of behavior, medical records, and neuropsychological test performance.

Slick et al. (1999) defined MND as "the volitional exaggeration or fabrication of cognitive dysfunction for the purpose of obtaining substantial material gain, or avoiding or escaping formal duty or responsibility. Substantial material gain includes money, goods, or services of nontrivial value (e.g., financial compensation for personal injury) and formal *duties* are defined as actions that people are legally obligated to perform (e.g., prison, military, or public service, or child support payments or other financial obligations). Formal *responsibilities* are those that involve accountability or liability within legal proceedings (e.g., competency to stand trial)" (p. 552). Three levels of classification are available in the Slick et al. criteria including *definite*, *probable*, and *possible malingering* with objectively defined criteria for each level. Common across all three levels of classification is the presence of a substantial external incentive, and that the presentation is not fully accounted for by other factors such as psychiatric, neurological, and/or developmental causes. Further distinction however is based upon the level of certainty and the nature and amount of clinical data suggestive of malingering.

Definite malingering is characterized by "the presence of clear and compelling data indicating volitional exaggeration or fabrication of cognitive dysfunction without plausible evidence of alternative explanations for such a cognitive profile" (p. 552). *Probable malingering* on the other hand is characterized by "the presence of evidence strongly suggesting volitional exaggeration or fabrication of cognitive dysfunction without plausible explanation" (p. 552). The major distinction between definite and probable malingering is that there is no clear evidence of a definite negative response bias for probable malingering. The lowest level of certainty in the Slick et al. criteria is *Possible malingering* which is essentially considered if evidence exists that to an extent suggests symptom exaggeration but the evidence is less strong than in probable or definite malingering. Alternatively, possible malingering is considered in cases where criteria for definite or probable malingering are met, but alternative explanations cannot be ruled out. Table 1 presents each of the diagnostic criteria as well as the required criteria for each level of diagnosis.

The estimated base rates of general malingering vary based on the nature of the referral question, the setting of the evaluation, and the classification scheme (Mittenberg, Patton, Canyock, & Condit, 2002; Russeler, Brett, Klaue, Sailer, & Munte, 2008), but recent surveys suggest that the base rate of malingering approximates 40% across multiple settings (Larrabee, Millis, & Meyers, 2009). At first glance, they can be quite staggering, with some estimates exceeding 85% in mild head injury cases when the estimated prevalence of persisting deficits in mild TBI is taken into consideration (Larrabee, 2007). A review of 11 studies published between 1978 and 2002 found the average estimated prevalence of malingering to approximate 40% in compensation-seeking mild head injury claimants when using the Slick et al. criteria (Larrabee, 2003), and a study of personal injury litigants and workers compensation claimants, found that the base rate of insufficient effort exceeded 40%, and effort accounted for up to 35% of the variance in neuropsychological test performance (Stevens, Friedel, Mehren, & Merten, 2008).

By surveying members of the American Board of Clinical Neuropsychology practicing in multiple clinical settings including litigating and non-litigating civil and medical settings, as well as criminal settings, Mittenberg et al. (2002) identified the highest rates of malingering among medico-legal cases in rehabilitation, medical, veterans, and psychiatric hospitals (31.0%) with private practice coming in a close second (29.8%); removal of the influence of litigation or compensation reduced these prevalence estimates to 11.6% and 7.1% respectively, demonstrating the powerful effect financial compensation has on test performance. Evaluation of prevalence estimates based on diagnosis reveals the three highest incident rates are among those reporting mild head injury (38.5%), fibromyalgia/chronic fatigue (34.7%), and pain/somatoform disorders (31.4%). Interestingly, only 8.8% of moderate to severe head injury cases were identified as malingering, the second lowest of all populations studied. These estimates reported by Mittenberg et al. are based on evaluation of multiple sources of objective evidence, the most common being a discrepancy between the severity of impairment in relation to injury severity and a pattern of performance on neuropsychological tests that was inconsistent with injury severity. Clearly, the provision of suboptimal effort and/or symptom exaggeration within the context of neuropsychological assessment is a serious and relatively commonplace issue.

The strategies employed by malingerers vary considerably in approach, efficacy, and sophistication; some are very easy to detect whereas others are much more deceptive. In one of the only qualitative post-hoc reviews of malingering research, Iverson (1995) surveyed participants from an analog malingering study and found that most laypeople have inaccurate beliefs about the sequelae of TBI. The most commonly employed approach to feigning memory impairment was to portray severe memory loss and total amnesia. Other commonly employed strategies were to fake partial amnesia and to “go blank” during the assessment. Iverson asked participants not only what strategies they employed during the simulation, but also asked what other ways people could think of to fake memory impairment. Interestingly, 30% of survey respondents could not identify any other way to fake memory dysfunction.

More recently, Axelrod (2008) found that the memory profile of individuals failing a combination of three symptom validity tests (SVT's) was significantly lower than the remainder of their cognitive profile, suggesting a deliberate attempt to falsify performance on tests specifically related to memory. Although those passing the effort measures also demonstrated memory impairment in relation to IQ functioning, those who failed the SVT's produced significantly worse memory profiles, even in comparison to those who passed the SVT's. These findings taken together further reinforce that of the major cognitive domains typically assessed in a thorough neurocognitive evaluation, memory is highly susceptible to non-credible performance and malingering.

Although clinicians ask participants to put forth full effort during the evaluation, many individuals may be tempted to put forth less than optimal effort for a myriad of reasons including financial incentive (Flaro, Green, & Robertson, 2007), mitigation of responsibility (Larrabee, 2007), fatigue (Majer, Welberg, Capuron, Miller, Pagnoni, & Reeves, 2008), or viewing the assessment as a tool being used to remove certain social freedoms (Rogers, 1997a). Although the latter may seem counterintuitive, some examinees may feel that they are powerless in the assessment process and that regardless of their performance they will lose certain aspects of their independence. As such, there is a prominent need for empirically-derived prediction models to aid in the detection of incomplete effort, as self-report based assessment of effort is too subjective and unreliable.

Section 1.4 – Assessment of Effort

In response to the increased risk for suboptimal effort, clinicians have developed strategies to detect client engagement in negative impression management strategies. Although clinical judgment is one tool towards that goal, previous research has demonstrated that even the most experienced clinicians are incapable of reliably and accurately detecting even simulated TBI in the absence of specifically designed effort measures (Heaton, Smith, Lehman, & Vogt, 1978). Within the malingering literature, there are several approaches to the detection of suboptimal effort, including evaluation of inconsistencies in the client's report of symptoms and/or test performances with other sources of clinical data (e.g., observation, collateral report, etc.), and performance on measures designed to indicate the provision of suboptimal effort; such measures are frequently referred to as "Symptom Validity Tests."

At present, a number of Symptom Validity Test's (SVT) are available to the clinical neuropsychologist for the purpose of evaluating the legitimacy of an examinee's cognitive profile, and their use is endorsed by national organizations including the National Academy of Neuropsychology (Bush, Ruff, Troster, Barth, Koffler, Pliskin, Reynolds, & Silver, 2005) and the American Academy of Clinical Neuropsychology (Heilbronner, Sweet, Morgan, Larrabee, Millis, & Conference, 2009). Independent, stand-alone symptom validity tests such as the Test of Memory Malingering (TOMM; Tombaugh, 1996), and Paul Green's verbal and non-verbal Medical Symptom Validity Test (MSVT; 2005) are commonly included within a neuropsychological battery. Although these measures have been developed by independent parties, they have two primary factors in common: first, they are both related to aspects of memory functioning, because memory is a cognitive domain highly susceptible to impression management strategies and dissimulation (Binder & Rohling, 1996; Suhr & Barrash, 2007; Williams, 1998). Second, these measures are constructed within a framework of a forced-choice paradigm (Hiscock & Hiscock, 1989), which is designed to detect below-chance performance based on statistically derived cut-off scores. Unfortunately, forced-choice validity measures are not very robust to even simple coaching (Suhr & Gunstad, 2007), and information about how to dissimulate successfully is sufficiently available via resources such as the internet for the commonly used forced-choice measures to allow examinees to avoid detection (Bauer & McCaffrey, 2006).

As a result, several “embedded” measures of effort have also been developed. These are typically composed of a specific index or score, from a standard measure (e.g., the digit span subtest of the Wechsler Adult Intelligence Scale, 3rd Edition ; Axelrod et al., 2006; Wechsler, 1997), or statistical prediction models that employ multiple scores from a particular measures to determine the probability of malingering, such as those developed for the California Verbal Learning Test (Coleman, Rapport, Millis, Ricker, & Farchione, 1998; Millis, Putnam, Adams, & Ricker, 1995; Wolfe, Millis, Hanks, Fichtenberg, Larrabee, & Sweet, 2010), and Repeatable Battery for the Assessment of Neuropsychological Status (Silverberg, Wertheimer, & Fichtenberg, 2007).

Similar prediction models have also been generated for the previous editions of the Wechsler Memory Scales (Iverson, Slick, & Franzen, 2000; Langeluddecke & Lucas, 2003; Ord et al., 2008). Sensitivity estimates of the primary index scores were highest for recognition scores, which ranged from 64% to 81% with 100% specificity

(Langeluddecke & Lucas, 2003). In a similar study investigating both individual index scores and aggregated summary WMS-III scores among mild TBI survivors, sensitivity ranged from 58% to 61% at 100% specificity (Ord et al., 2008). Although the specificities of these prediction methods are quite impressive, the sensitivities are too low to be considered useful in isolation from other effort measures. They do, however, represent a viable option to be used in conjunction with other diagnostic tools, and when combined with failure of another SVT, the probability of poor effort approaches 100% (Larrabee, 2008)

In contrast to stand-alone effort measures, embedded measures typically are not readily identifiable as measures of effort, making it is less likely that an examinee will alter his or her response style for that subtest and put forth high effort in order to avoid detection. Statistically based prediction models present a very difficult scenario for examinees to dissimulate successfully, as doing so would require a sophisticated understanding of the measure and a very complex performance strategy (Suhr & Gunstad, 2007). Embedded measures and prediction modes have the added benefit of providing useful clinical data regarding cognitive ability (only when poor effort is not a concern), thus increasing efficiency and maximizing the utility of information garnered in a standard assessment. In addition, developing measures of effort embedded within standard neuropsychological measures affords clinicians to investigate the credibility of a cognitive profile when data from independent SVT’s is unavailable such as may be the case in forensic record reviews (van Gorp, Humphrey, Kalechstein, Brumm, McMullen, Stoddard, & Pachana, 1999). Hence, the best practice is to employ a combination of both stand-alone and embedded symptom validity tests, the findings of which can be considered together to formulate clinical decisions regarding the validity of an examinee's test performance (Greve, Ord, Curtis, Bianchini, & Brennan, 2008).

Section 1.5 – Coaching and Effort Testing Research

Effort testing research is a sensitive subject of study, as balancing test security, development of detection strategies, and publication of findings can be difficult. There are a number of ethical issues that come into play within this area that are not necessarily prevalent in other areas of clinical research. With the rapid increase of empirical literature on effort testing and psychological measures over the past decade and the availability of publicly-accessible information (Bauer & McCaffrey, 2006; Ruiz, Drake, Glass, Marcotte, & van Gorp, 2002) comes increased uncertainty as to how the data will be used and by who (Iverson, 2006). Although the research is intended to aid practicing clinicians and promote further research, there is little stopping a patient or the savvy attorney from accessing such information and using it to their advantage. In a survey of members of the American Trial Lawyers association, it was revealed that many attorneys spend up to an hour coaching their clients prior to a neuropsychological evaluation (Essig, Mittenberg, Petersen, Strauman, & Cooper, 2001). One of the tools likely used are the empirical articles on malingering reporting detection abilities of various measures. Given this potential, many researchers refrain from publishing sufficient information (e.g., logistic regression variables) in their manuscripts that would allow for detecting the presence of an SVT and subsequently altering their response patterns. Instead, they provide the information upon request in an attempt to keep better track of who is using the data and how.

Coaching – the provision of sufficient information intended to help an examinee portray themselves in a disingenuous way (Suhr & Gunstad, 2007) – is still highly prevalent and can take multiple forms. In addition to the internet and attorneys, repeated neuropsychological evaluations which are common in a forensic case can provide information to clients on how to take cognitive tests, especially when feedback on performance is given (Suhr & Gunstad, 2007). Given the prominence of coaching within the forensic neuropsychology arena, this area of practice lends itself well to research with the aims of developing better detection strategies that are more robust to coaching (Rogers, 1997b). The three most common methods of coaching within the context of a neuropsychological evaluation include the provision of information related to injury symptoms, specific strategies on how to take tests (e.g., “if you see a test where you have to choose between two pictures, do your best”, etc.), and warning clients of the use of SVT’s within the assessment (Suhr & Gunstad, 2007). If neuropsychological assessment is to remain a valid and accurate practice, effort testing research is a critical and essential endeavor toward its preservation.

Several approaches to effort testing research exist, which include case study, differential prevalence designs, simulation studies, and known-group designs (Larrabee, 2007; Rogers, 1997b). The weakest and least frequently used is the differential prevalence design in which a researcher estimates the prevalence of malingering and non-malingering in given sample (Rogers, 1997b). Unfortunately, however, there is little empirical evidence supporting the classification of group membership and such discriminations are typically inferred based on existing literature as opposed to being measured directly (Larrabee, 2007). The case study approach clearly has extremely limited generalizability and as such is rarely, if ever, used in clinical research. It is, however, an essential aspect of clinical work and individual case conceptualization as examining individual patterns of performance within a cognitive profile can lead to hypothesis generation and further investigation on a case by case basis (Larrabee, 2007).

The most commonly used malingering research approach to date has been the simulation study (Suhr & Gunstad, 2007); however, the use of known-group designs is on the rise based on criticisms of analog designs (Greiffenstein et al., 1995; Larrabee, 2007). In a simulation study, normal participants are coached in some way to feign cognitive impairment. The major issue with an analog simulator group is that there is little, if any, assurance that the simulators will perform in the same way that true clinical malingering patients would, which limits generalizability and applicability in a forensic arena (Larrabee, 2007). Although there are ways to improve the generalizability of findings from analog studies (e.g., randomly assigning clinical patients to a simulation or control condition; Rogers, 1997b), the feasibility of such research is questionable and the ethics of randomly assigning bona-fide TBI patients involved in real world assessments to malingering or honest effort conditions is highly debatable.

Perhaps the most robust research approach is the known-groups design which utilizes specific objective criteria to classify clinical patients into two or more groups based on objectively defined criteria and makes comparisons between the groups (Larrabee, 2007; Rogers, 1997b). The greatest advantage of the known-group design is that it dramatically increases the generalizability of findings as the sample is drawn from real world patients and the feigning of cognitive impairment is occurring *in vivo* (ibid.).

The two major drawbacks to this methodology are that random assignment is lost entirely as classifications are based on objective criteria, typically after completion of the assessment, and classification itself can be difficult (Larrabee, 2007). With the advent of objectively defined classification criteria such as those developed by Slick et al. (1999), group classification has been greatly improved and has thus facilitated the use of known-groups designs. Unfortunately however, there is little that can be ethically done to address the issue of random assignment.

Although each approach to researching malingering has its drawbacks, each also has its place. In the early phases of SVT development and malingering detection strategies, analog studies are well suited as there is minimal risk involved to participants and generalizability is of less importance than test development. Once a measure has been developed and proven using an analog study, further validation using a known-group design is warranted and the appropriate follow-up to the analog study. Alternatively, a blended design can also be an excellent approach to aid in both the development and validation of a malingering detection strategy. For example, by using a three-group design including simulators, full-effort clinical patients, and known malingerers using post-hoc classification, a single study can both develop a detection method by comparing performance between the simulator group and honest-effort group, which can then be validated in comparisons between the honest-effort and poor effort group.

Section 1.6 – Efficacy and Outcome Research

Evidence-based medicine is standard practice within the fields of mental health (APA, 2006) and general medicine (AMA, 1992), and recently it has been indicated for use in rehabilitation whenever possible (DenBoer & Hall, 2007; Tate, Kalpakjian, & Kwon, 2008). Research into the efficacy of intervention programs within the rehabilitation setting is critical to the advancement of the field and to the greater well being of the population being served. By understanding what works in the rehabilitation environment, we are better able to effectively treat those in need. A necessary factor in understanding the efficacy of specific interventions is the randomized controlled trial, which investigates efficacy of an intervention in comparison to a gold standard form of treatment using reliable outcome measures (Piantadosi, 2005). As such, the results of these studies are highly dependent on the measures used as well as the level of effort put forth by participants on those outcome measures. For example, poor effort may artificially attenuate efficacy estimates, making an effective intervention appear ineffective or less effective than it truly may be. By developing empirically-validated, embedded measures of effort within a cognitive assessment battery such as the WMS-IV that can be used in conjunction with existing stand-alone effort measures, we can effectively screen for effort and account for individuals in analyses and subsequent participation that have provided less than optimal effort.

Section 1.7 – Limitations of the extant literature

Although there is a substantial amount of peer-reviewed literature investigating the assessment of effort put forth during cognitive evaluations, the WMS-IV is essentially an all-new measure and has been almost entirely revised from its previous version. As such, there is a dearth of research on this measure except for studies conducted by the test developers that are included in the technical manual; most frequently, studies of effort assessment are left to the realm of independent clinical researchers. Although clinicians would be able to utilize the previous version of the Wechsler Memory Scale in order to include validated embedded measures of effort, using obsolete instruments would not be considered an acceptable standard of care (APA, 2002) and it would quickly become inconsistent with state-of-science expectations for the assessment of memory function.

Additionally, it remains possible for very sophisticated dissimulators to slip past even well developed measures of effort. With essentially only a small handful of qualitative studies on malingering strategies, what remains unclear is how successful malingerers are succeeding. More specifically, it is presently unknown what strategies are employed to avoid detection on effort measures by persons who successfully feign impairment. Gaining better perspective of the strategies used will aid in the development of future effort measures that are robust to these strategies.

CHAPTER 2

AIMS OF THE CURRENT STUDY

The present study sought to add to the literature regarding use of the WMS-IV among persons with TBI by investigating its ability to differentiate cases of bona fide TBI from feigned memory impairment. Similar to the work of Ord et al. (2008) who used the WMS-III, the development of empirically-based strategies for the detection of insufficient effort using the WMS-IV was of substantial interest, especially considering the significant impact effort has on neuropsychological test performance (Stevens et al., 2008) the susceptibility of memory tests to feigned cognitive dysfunction (Binder & Rohling, 1996), and the common use of Wechsler family tests in assessment practices (Rabin et al., 2005). Not only are data obtained from neuropsychological assessments used for the development of therapeutic interventions for TBI survivors (Bush et al., 2005; Pereira, 2007; Stringer & Nadolne, 2000), they are also a common outcome measure in efficacy studies investigating specific interventions.

Section 2.1 – Specific Objectives

A. *Determine the ability of the WMS-IV and its subtests to reliably distinguish between cases of dissimulated brain injury (SIM) and verified traumatic brain injury (TBI).*

A primary objective in the present design was to identify variables within the WMS-IV that contribute to the identification of persons putting forth suboptimal effort during their evaluations. In addition, this study sought to determine the decisional accuracy of generated prediction models in comparison to existing methods of effort assessment.

Hypothesis 1: It was predicted that individual indices or a combination of indices obtained from administration of the WMS-IV would be capable of reliable differentiation between persons with verified TBI and neurologically-normal participants dissimulating TBI.

B. *Determine the ability of the WMS-IV and its subtests to reliably distinguish between cases of successfully dissimulated TBI and unsuccessfully dissimulated TBI.*

A second objective in the present design was to determine the pattern of performance that characterized individuals asked to dissimulate TBI who pass independent measures of effort.

Hypothesis 2: It was predicted that individual indices or a combination of indices obtained from administration of the WMS-IV would be capable of reliably differentiating between successful TBI dissimulators and unsuccessful TBI dissimulators.

CHAPTER 3

METHODS

Section 3.1 Participants

Data for the present study were collected from persons in the Detroit Metropolitan area and surround suburbs. Two primary samples were collected including persons with traumatic brain injury (TBI) and healthy adults coached to feign cognitive impairment (SIM); these samples were subsequently further divided based on the number of effort tests failed. Traumatic brain injury survivors were recruited from the pool of individuals presently enrolled in the Southeastern Michigan Traumatic Brain Injury System (SEMTBIS) research project who agreed to be contacted regarding future research projects. As a result of inclusion criteria, this sample excluded persons with mild injuries or very severe brain injuries who did not receive inpatient rehabilitation. A full listing of inclusion criteria for the SEMTBIS project can be found in Appendix B. The initial sample size recruited and tested for the TBI group was 61 people. After completion of testing, persons in this group were classified as either putting forth good or poor effort based on the number of effort measures failed; there was no consideration given to which two effort measures were failed. Failure of two or more was considered poor effort based on previous research suggesting that with a base rate of approximately 40%, failure of two or more effort measures yields a 94% probability of poor effort (Larrabee, 2008).

A demographically comparable group of healthy adults coached to dissimulate TBI was recruited from the greater Southeastern Michigan region via newspaper advertisements and posted fliers. Exclusion criteria for this sample were such that the presence or history of neurological dysfunction or injury such as seizure, brain injury, stroke, or concussion prevented participation. Attempts were made to match this sample on key demographic factors including age and education. The initial pool of healthy individuals that completed testing included 65 people. Just as with the TBI sample, failure of two or more effort measures was used as the criterion for classification as putting forth poor effort. In addition, anyone who reported not following instructions on the debriefing questionnaire (e.g., indicating that they did not try to fake a brain injury) was removed from analysis as this report gave indication that the participant failed to follow instructions.

Section 3.2 - Measures

Section 3.2.1 - Demographic and Health Behavior Factors

In order to account for variance in outcome that is attributable to demographic and health status, information about gender, age, race/ethnicity, years of education, comorbid illnesses, and medication usage was recorded.

Section 3.2.2 - Injury Severity

Time to follow commands, which is defined as the number of days that it takes to obtain a score of 6 on the motor subscale of the GCS two out of two times within a 24-hour period, was used as an index of brain injury severity (Dikmen et al., 1995; Rohling, Meyers, & Millis, 2003). This value along with other indices of injury severity (e.g., length of post-traumatic confusion, initial GCS score) was used to account for variance in outcome that is attributable to TBI severity. Participants with GCS scores of 9 or greater were classified as mild/moderate injuries based. Traditionally, a GCS score of 13 to 15 is considered a mild injury; however, in the present sample the inclusion criteria for the SEMTBIS project were such that participants with GCS scores in the mild range also had documented intracranial hemorrhage significant enough to warrant inpatient rehabilitation. Empirical findings indicate that the cognitive profiles of persons with complicated mild injuries (e.g., a mild TBI with an intracranial bleed) are more similar to persons who have moderate brain injuries than to those who have uncomplicated mild brain injuries (Kashluba et al., 2008). Participants with a GCS score of less than 9 were classified as severe.

Section 3.2.3 - Memory Functioning

Wechsler Memory Scale – 4th Edition (WMS-IV): The primary measure of memory functioning was the WMS-IV, which is the latest revision of the measure released in January of 2009. The WMS-IV is a battery of subtests designed to evaluate multiple forms of memory functioning including immediate and delayed memory, as well as working memory, visual memory, and auditory memory. Complete administration of the WMS-IV generates 5 individual index scores, 12 primary subtest scores, 9 secondary subtest scores and 13 contrast scores. Raw scores as well as age-adjusted scaled scores from this measure were evaluated for use as predictor variables in the primary logistic regression.

California Verbal Learning Test – 2nd Edition (CVLT-II)(Delis, Kramer, Kaplan, & Ober, 2000): A secondary measure of memory functioning that was used as a gold-standard comparison for the WMS-IV is the CVLT-II. This measure is a verbal list-learning task that requires test takers to learn a list of 16 words from four semantic categories presented verbally over a series of five consecutive learning trials. Subsequent to the initial learning trials, a distracter set is presented followed by a short delay free recall of the initial learned set of words. Following a 30-minute delay, a second free recall portion is administered to assess long-term retention. In addition to the free recall trials, there are two separate cued recall trials and a forced choice recognition trial.

In the standard administration procedures of the WMS-IV outlined by the test publisher, the Trials 1-5 T score from the CVLT-II can be substituted for Verbal Paired Associates 1 and the Long-Delay Free Recall Z score can be substituted for Verbal Paired Associates 2. According to the test publisher, substitution of the CVLT-II scores generates scaled scores that correlate very highly with those obtained from standard administration. Although the publisher strongly discourages administration of both tests to avoid intra-individual comparison of substituted scores to non-substituted scores, in the present design both tests were administered and such comparisons were made in order to determine both the validity of this substitution as well as to determine the most discriminating variable set.

Section 3.2.4 – Effort

In addition to the primary measures of learning and memory, a total of seven effort measures were administered to each participant including four stand-alone measures and three embedded measures. With one exception (Word Choice Test; described below), performance on the individual effort measures was classified as either passing or failing based on published criteria. As previously described, the number of failed effort measures was then used to classify participants as either good (failed fewer than two) or poor (failed more than one) effort.

Test of Memory Malingering (TOMM): The TOMM (Tombaugh, 1996) is a 50-item, forced-choice visual recognition measure that is commonly used to assess effort.

Although the TOMM has demonstrated good specificity in detecting insufficient effort (Gierok & Dickson, 2000; Rees, Tombaugh, Gansler, & Moczynski, 1998; Teichner, Wagner, & Newman, 2000), recent research has failed to produce adequate sensitivity to be used in isolation from other effort measures (Greve et al., 2008). As such, it is typically incorporated into psychological assessment batteries along with other effort measures. According to the test manual, a raw score of 45 or less on Trial 2 is considered a failure; data from Trial 1 are not used and the optional retention trial was not administered.

Medical Symptom Validity Test (MSVT): The MSVT is a shortened version of the Word Memory Test. It is a forced-choice, computer administered measure of verbal memory that contains embedded measures designed to evaluate the validity of an examinee's presentation on the test (P. Green, 2005). Multiple studies have demonstrated the ability of the MSVT to detect suboptimal effort (Carone & Turk, 2008; Merten, Green, Henry, Blaskewitz, & Brockhaus, 2005; Richman, Green, Gervais, Flaro, Merten, Brockhaus, & Ranks, 2006). Data obtained from administration of this measure were used to classify individuals as putting forth good or poor effort. According to the manual, a score of 85% or less on any of the immediate recall, delayed recall or consistency measures is considered a failure.

Non-Verbal Medical Symptom Validity Test (NV-MSVT): In addition to the MSVT, a similar measure was also administered that used pictures of item pairs (e.g., a dog with a bone) as opposed to word pairs. The non-verbal medical symptom validity test (NV-MSVT; P. Green, 2008) may be more difficult than the MSVT because it requires a shift in target recognition such that unbeknownst to the test-taker, foils later become targets;

however, similar scores are generated including percentage correct for a number of trials including immediate, delayed, and cued recall. The failure criteria for this measure are a bit more complex than the verbal version of this measure. According to the manual, if the mean of the immediate recall, delayed recall, consistency, the two delayed recall conditions (archetypes or variations) and paired associates is below 90% *or* the mean of the two delayed recall conditions, overall delayed recall, and consistency is below 88%, the test is considered failed.

Wechsler Adult Intelligence Scale – 4th Ed. Reliable Digit Span (RDS): Calculation of the reliable digit span is frequently used as an embedded measure of effort (Greiffenstein et al., 1995). The RDS is calculated by summing the item span of the last trial where both items were correctly recalled. Although the original criterion for poor effort was considered a RDS of 7 or less, some have argued that this often results in an unacceptable false positive rate in cases of severe injury (Babikian et al., 2006). As a result, in the present study, the criterion for poor effort was lowered to a RDS of 6 or less.

Embedded Indices within the CVLT-II: Within the CVLT-II, there are two embedded effort measures. The first is a forced-choice recognition trial developed by the publisher and included at the end of administration. As with many of the stand-alone effort measures – particularly one with so few items – the sensitivity to brain injury is very low with high specificity. For the present study, the criterion for poor-effort was set at a score of 14 or less. In addition to the forced choice measure, a prediction equation has been validated using three indices obtained from standard administration (Wolfe et al., 2010) that generates a probability of poor effort.

Pearson Advanced Clinical Solutions: Following the release of the WMS-IV, Pearson released the Advanced Clinical Solutions (ACS) that contained five embedded effort measures for use with the WMS-IV. One of these measures is the Word Choice Test *(WC),* a 50-item forced-choice measure in which participants are presented a list of targets and later asked to identify the target from a single foil. Of all the effort measures administered, this was the only measure not included in the classification of participants as either good or poor effort, in order to avoid contamination of criterion with predictor. The four additional embedded measures include the Reliable Digit Span from the WAIS-IV, Visual Reproduction Recognition, Logical Memory Recognition, and Verbal Paired Associates Recognition. The ACS does not provide cutoff scores for classification of poor effort, but rather provides comparisons of performance to the overall clinical samples used in the validation process of the WMS-IV and WAIS-IV in the form of base rates.

Section 3.2.5 - Pre-morbid Intelligence

Wechsler Test of Adult Reading (WTAR): The WTAR is a single word-reading test of increasing difficulty that generates an estimate of overall intellectual ability (full scale IQ). Existing research on the WTAR has found it to be a valid estimate of intellectual functioning regardless of the presence of TBI (R. Green, Melo, Christensen, Ngo, Monette, & Bradbury, 2008). An estimate of premorbid intelligence also was calculated to account for variance in outcome measures that may be attributed to overall intellectual ability.

Section 3.3 - Procedure

Section 3.3.1 - Traumatic Brain Injury group (TBI)

Persons with TBI (n = 61) were contacted via telephone and informed of the opportunity for additional voluntary participation in a research study investigating use of a new memory assessment tool. All participants who agreed to participate were evaluated at the Rehabilitation Institute of Michigan's main campus following the provision of informed consent. Testing was completed in a single session by the primary investigator or a research assistant and participants were compensated $30 for their time upon completion of the study. Each TBI participant was asked to put forth their full effort on all measures attempted. The order of administration varied among participants due to differences in test performance; some participants completed the measures intended to occupy delay periods quickly, whereas others took longer to complete such measures. The administration of either the CVLT-II or the WMS-IV as the first memory measure, however, was counterbalanced across the TBI sample to reduce order effects. Following administration, all measures were scored according to standardized procedures. For this group, persons with TBI were classified as good effort if they obtained passing scores on at least 5 out of 6 effort measures.

Section 3.3.2 - Dissimulated Traumatic Brain Injury group (SIM)

Participants in the coached group (n = 65) were neurologically normal persons recruited from the Detroit Metro area community via printed advertisement (e.g., Metro Times, Detroit News, etc.) as well as the staff and students of Wayne State University. Participants were excluded from the study if they reported a history of neurological condition or event (e.g., seizure disorder, TBI, etc.). As with the TBI group, testing took place at either the Rehabilitation Institute of Michigan (RIM) or the research laboratory of the primary investigator.

Following provision of informed consent, participants in the SIM group were administered the WTAR under instructions to put forth full effort to obtain an accurate estimate of intellectual ability. After completion of the WTAR, participants were then told that a particular focus of the remainder of the study was to determine the ability of a new test of learning and memory to assess the amount of effort put forth during a cognitive examination. Each participant was then presented with a scenario indicating their involvement in litigation following a motor vehicle accident that resulted in a TBI. The scenario described was constant for all SIM participants and was read from a script used successfully in prior research on dissimulation of this basic design (DenBoer & Hall, 2007; Tombaugh, 1997). A complete copy of the script read to each participant, is presented in Appendix A.

Consistent with prior research on dissimulation of TBI and based on the recommendations outlined by Suhr and Gunstad (2007), participants in the SIM group were also provided with literature describing the nature of the cognitive impairment typically resulting from TBI (e.g., memory dysfunction, reduced processing speed, etc.; Coleman et al., 1998). Participants were given time to read the literature and then asked to restate both the instructional set and the common symptoms of TBI in their own words to ensure comprehension; if a participant was unable or unwilling to follow the instructions, they were compensated for their time and excused from participation. Each SIM participant was evaluated in a single session. As with the TBI group, the order of administration was variable in order to fulfill required delay periods, with the administration of the memory measures counterbalanced across the SIM sample.

Section 3.3.3 - Debriefing

Following the completion of the assessment battery, all participants in the SIM group completed a questionnaire asking them to provide their strategies for dissimulation. More specifically, each participant first reported whether or not they followed instructions to feign cognitive impairment (an indication of not following instructions removed the participant’s data from analysis). If they reported following instructions, they were then asked to rate the level of difficulty on a 7-point Likert scale ranging from “Very Easy” to “Very Hard”. Participants were also given the opportunity to provide a written explanation of the strategies employed during the assessment.

CHAPTER 4

ANALYSES

Section 4.1 - Data Coding

All cases in the TBI group were screened for the provision of poor effort: To remain eligible for inclusion in subsequent analyses, at least five of the six data validity indicators (e.g., MSVT, TOMM) were negative for the provision of suboptimal effort. This cutoff was chosen based on previous research (e.g., Larrabee, 2008) and the assumption that few clinicians would consider a participant to be globally putting forth poor effort based on a single failed effort measure. As a result of this classification procedure, the remaining sample of TBI survivors can be considered to have provided good effort and generated valid results. This sample was further divided based on injury severity using the GCS score taken at the time of admission to the emergency department into a mild/moderate TBI group and severe TBI group.

For the simulator group, in order to be included in further analysis, participants had to have reported feigning cognitive impairment on the debriefing form. If they did not, their data were removed from subsequent analysis, as it was assumed they failed to follow instructions. From this point, data coding in the SIM group was similar to the TBI group and produced two groups based on variability in presentation. The primary simulator group of interest was those individuals who failed two or more of the symptom validity tests; this group is referred to as "unsuccessful simulators" using the logic that they were not successful in avoiding detection on formal effort testing. Those who failed one or none of the effort measures were coded as "successful simulators" in the sense that they were successful in avoiding detection.

Given the possibility that simulators classified as "successful" may not have followed instructions despite stating they had done so on the debriefing form, this sample was further divided into "impaired" or "intact" based on their WMS-IV performance. In order to be considered "impaired", the participant had to have obtained at least one index score that was 1 standard deviation or more below the mean (e.g., ≤ 85).

Section 4.2 - Primary Analyses

The primary analytic strategy for the present study (e.g., binary logistic regression and ROC curve analysis) has been successfully employed in similar research conducted on data collected from our outpatient rehabilitation population (Miller, Millis, & Fichtenberg, 2010). In the present study, binary logistic regression was used to compare unsuccessful simulators to good-effort TBI. Mean values were calculated for each of the primary index scores as well as the subtests and secondary measures. Using a pooled standard deviation, effect sizes (Cohen's *d*) were calculated to aid in selection of viable predictors. Those variables demonstrating large effect sizes were considered candidates for further analysis. Exploratory models were built using those WMS-IV scores identified as potential predictors and group membership (TBI vs. SIM) as the outcome. This process was repeated using several combinations of secondary measures as well as the subtest scores.

From these initial models, collinearity diagnostics were examined in order to determine the potential for statistical overlap between variables, and odds ratios were examined to determine the relative influence of each variable on identification of group membership. In the presence of significant overlap between two or more variables based on collinearity diagnostics, a theoretically guided decision based on both observed odds ratios and relevant existing literature was made as to which variable(s) should be removed from subsequent models. This process was repeated until a model was produced that was void of substantial collinearity and represented the most parsimonious model. Because index scores are composed of multiple measures, however, this approach was too broad to identify the most discriminating variables and subtests. As a result, this analytic approach was repeated using individual subtest scores.

Following generation of models composed of only unique variables, the models were compared using Bayesian information criterion (BIC) statistics to determine which model best fit the observed data. Receiver operating characteristic (ROC) curves were also generated to aid in determination of classification hit rate, and sensitivity and specificity of each predictive model. Positive and negative predictive powers were not calculated due to the artificial base rate created by the study design.

Section 4.3 - Secondary Analyses

Following completion of the primary analyses, secondary analyses were conducted on the simulator groups to determine if a performance pattern exists that distinguished successful from unsuccessful dissimulators. An identical analytic approach to the primary analyses (e.g., binary logistic regression and ROC curves) was employed using SIM status (successful vs. unsuccessful) as the outcome variable.

CHAPTER 5

RESULTS

Section 5.1 – Sample Demographics

The overall TBI group (*n* = 60) was predominantly African American (81.7%) men (93.3%) with a mean age of 43.9 years (*SD* = 12.1), mean education of 12.0 years (*SD* = 2.2). The overall SIM group (*n* = 64) was also predominantly African American (68.8%) men (82.8%) with a mean age of 43.9 years (*SD* = 11.4), mean education of 12.9 years (*SD* = 2.1). Following classification of effort level using the previously described methods, four subgroups were obtained, which are the primary groups of interest to be used in subsequent comparisons: 1) Good Effort TBI (*n* = 41); 2) Suspect Effort TBI (*n* = 19); 3) Successful Simulators (*n* = 18); and 4) Unsuccessful Simulators (*n* = 41). Five participants were removed from the simulator group, as they did not follow coaching instructions, and one was removed because of excessive missing data. Demographic data for each subgroup are presented in Table 2. The four effort groups were compared on all demographic factors and no significant differences were observed for age (F (3, 115) = 2.13, p = .10), education (F (3, 115) = 2.46, p = .066), gender (χ^2 (3) = 5.09, p = .165) or ethnicity (χ^2 (2) = 8.09, p = .232).

The Good Effort TBI group was 51.2% mild/moderate injuries (*n* = 21) and 46.3% severe injuries (*n* = 19); one case was missing injury severity data. Time since injury ranged from 10 to 234 months for Good Effort TBI participants, and the mean number of days in acute rehab was 26.5 days (*SD* = 13.9; range = 6.0 to 66.0). The average time to follow commands was 7.1 days (*SD* = 11.3; range = 0.5 to 41.0) and the average time to clear post-traumatic confusion was 22.4 days (*SD* = 19.5; range = 0.0 to 72.0).

The Suspect Effort TBI group was 52.6% mild/moderate injuries (*n* = 10) and 42.1% severe injuries (*n* = 8); one case was missing injury severity data. Time since injury ranged from 10 to 234 months for Suspect Effort TBI participants and the average length of stay in acute rehab was 23.7 days (*SD* = 14.4; range = 7.0 to 56.0). The average time to follow commands was 5.4 days (*SD* = 7.2; range = 0.5 to 25.0) and the average time to clear post-traumatic confusion was 22.8 days (*SD* = 18.0; range = 0.0; to 64.0). There were no significant differences found between the Good Effort and Suspect Effort TBI groups on any of the injury related variables.

<u>Section 5.2 – Test Performance</u>

Section 5.2.1 – Wechsler Memory Scale – Fourth Ed. (WMS-IV)

Performance on the WMS-IV for each of the four subgroups is summarized in Table 3 (Index Scores), Table 4 (Substitution Scores) and Table 5 (Secondary Scores). Figure 1 presents a subtest score profile for each of the four groups. Effect sizes (Cohen's *d*) are also presented in the tables and are based on a pooled standard deviation comparing the Good Effort TBI group to the Unsuccessful Simulator group in order to evaluate the effect of coaching in the present study. These effect sizes were also used to guide variable selection for entry into subsequent prediction equations.

Section 5.2.1.1 Good Effort TBI

One-way analyses of variance (ANOVA) compared the mild/moderate TBI and severe TBI groups on each of the primary index scores and did not reveal the presence of significant differences. As such, the two severity groups were collapsed and the WMS-IV performance of the combined overall Good Effort TBI group was used in subsequent analyses. For the Good Effort TBI sample, the majority of participants

(39.0%) demonstrated an impaired performance (defined as at least 1 SD or more lower than the mean) on all five index scores using the standard calculation; using the CVLT-II substitution method, 36.8% were impaired on all five index scores. Less than 10% of the sample scored within 1 SD of the mean on all five of the index scores with the standard administration, and 13.2% did so when substituting CVLT-II indices. As can be seen in Table 3, on four out of five standard index scores, over 60% of Good Effort TBI participants demonstrated impairment; the exception was the VMI, on which slightly less than half the sample showed impaired performance. Each of the three index scores based on CVLT-II substitution were impaired for 50% or more of the Good Effort TBI sample. In comparison to the Unsuccessful Simulator group, the Good Effort TBI group demonstrated a significantly lower proportion of impaired scores on the VMI (χ^2 (1, *N* = 82) = 9.02, *p* = .003, phi = .33) and DMI (χ^2 (1, N = 82) = 7.75, *p* = .005, phi = .31). The two groups did not differ significantly on the proportion of impaired scores for AMI, VWMI, or IMI (all *p*'s > .10). In comparison to the Successful Simulator group, the Good Effort TBI group demonstrated a significantly higher proportion of impaired scores on the AMI (χ^2 (1, *N* = 59) = 10.68, *p* = .001, phi = .43) and VWMI (χ^2 (1, *N* = 59) = 6.47, *p* = .011, phi = .33), whereas the groups did not differ significantly on VMI, IMI, and DMI (all *p* values > .10).

As can be seen in Figure 1, the Good Effort TBI group generally produced lower scores on verbal subtests in comparison to visually based subtests. Using the same criteria of impairment (e.g., 1 SD or more below the mean), subtest means were in the impaired range for 5 out of 10 primary subtests. For most subtests, more than 50% of the Good Effort TBI sample demonstrated impairment. The exceptions to this were on the delay portion of Designs (31.7% impaired), and both immediate (39.0%) and delay (48.8%) portions of Visual Reproduction.

Recognition measures are summarized in Table 5. Raw scores ranged from 11 to 29 (out of 30) for Logical Memory (LM) recognition and from 2 to 7 (out of 7) for Visual Reproduction (VR) recognition. For LM recognition, 73.2% of participants achieved at least 65% accuracy and on VR recognition, 68.3% of participants obtained a score of 6 or better. On the recognition portion of the Designs subtest, scores ranged from 8 to 18 out of 24 with 65% of participants correctly identifying 50% of more of the targets. For Verbal Paired Associates (VPA) recognition, in addition to a simple hit rate, the WMS-IV also calculates the number of Easy Hits, Hard Hits, False Positives, and a discriminability score. There are a total of 14 targets (4 of which are Easy) and 26 foils. Total hits ranged from 15 to a maximum of 40 with a median of 36; all but one Good Effort TBI participant achieved at least 67% accuracy. Easy hits ranged from 1 to 4 with 90.3% of the sample correctly identifying at least 3 of the easy items and 73.2% of the sample correctly identifying all 4. Hard hits totals ranged from 2 to 10 with 75.9% of the sample correctly identifying 9 (n = 16) or all 10 (n = 15) of the Hard items. False positives ranged from 0 to 16 with 30.0% of the sample not making any false positives errors and 65.0% making 3 or less false positive errors; only 2 people in the Good Effort TBI group made more than 10 false positive errors. Verbal Paired Associates recognition discriminability is calculated using the following formula:

$$\text{Discriminability} = (\text{Hits} / 14 + ([26 - \text{False Pos.}] / 26)) / 2$$

Values for the discriminability range from 0.0 to 1.0 and are considered an overall measure of recognition accuracy that weights false positives along with the total number of hits. In the good effort TBI sample, discriminability ranged from .37 to 1.00 with a mean of .89 (*SD* = .11) indicating that participants in this group were able to correctly separate out the correct items with relatively few false positives errors. Ninety percent of the sample achieved a discriminability score of .78 or better.

Although not included in the calculation of any index score, the Brief Cognitive Status Exam (BCSE) is essentially an expanded version of the Information and Orientation subtest from the WMS-III. The BCSE is divided up into seven sections including Orientation, Time Estimation, Mental Control, Clock Drawing, Incidental Recall, Inhibition, and Verbal Production. Scores from each section are weighted and summed, and the total score is classified based on age group and education as Average, Low Average, Borderline, Low, or Very Low. Good effort TBI participants generally performed well on this measure with raw scores ranging from 23 to 58 (out of 58) and 53.7% of participants being classified as Average, and 22.0% being classified as Low Average.

Section 5.2.1.2 Suspect Effort TBI

The suspect effort group consisted of outpatient adults who were part of the SEMTBIS project who failed two or more of the symptom validity tests; none reported involvement in current litigation or other legal proceedings. As with the Good Effort TBI group, one-way ANOVA’s compared each of the primary index scores by injury severity and did not reveal the presence of significant differences, so the suspect effort TBI group was also combined across injury severity. The majority of participants in the

Suspect Effort TBI group (57.9%) demonstrated impaired performance on all five of the standard index scores; using the CVLT-II substitution method, 52.6% were impaired on all five of the index scores. Using either method (CVLT-II substitution or standard administration), the entire Suspect Effort TBI sample was impaired on three or more of the index scores. For every subtest, including the substituted scores, over 60% of this group scored in the impaired range with some as high as 95% (Spatial Addition).

Comparing each of the index scores between the Good effort and Suspect effort TBI groups using one-way ANOVA revealed significant differences (all p values $< .005$; η^2 range from .19 to .22), thus confirming that those individuals in the TBI group who were classified as putting forth questionable effort based on SVT failure, performed significantly worse on the primary WMS-IV indices. A similar pattern of findings was revealed at the subtest level. More specifically, the Suspect Effort TBI group performed significantly worse on most subtests in both the immediate and delay portions with the exception of the Designs subtest, which approached significance. On the recognition measures, the Suspect Effort TBI group again performed significantly worse except for LM recognition where the observed difference was not significant.

Section 5.2.1.3 Unsuccessful SIM

Unsuccessful Simulators (U-SIM) represent the primary group of comparison against the Good Effort TBI that were used in generation of the primary prediction equations. All participants in this group were healthy adults who failed at least two of the six effort measures. On the WMS-IV, the number of impaired index scores ranged from 1 to 5 using the standard administration and from 0 to 5 using the CVLT-II substitution. With the standard administration, 70.7% of U-SIM participants demonstrated impairment on all five of the index scores using the standard administration; this value dropped to 65.9% with the substitution method.

In comparison to the Good Effort TBI group, the number of impaired index scores was significantly higher for U-SIM participants using either the standard administration (Mann-Whitney U = 561.5, p = .004) or the substitution method (Mann-Whitney U = 510.0, p = .004). Additionally, U-SIM participant performance did not differ significantly on the AMI, although a significant trend was observed (F (1, 80) = 3.736, p = .057, η^2 = .05). For all other index scores, including the substitution scores, the U-SIM participants performed significantly worse than the Good Effort TBI participants (p < .005). Evaluation of Cohen's *d* calculated using a pooled standard deviation approach revealed effects that ranged in size from medium (AMI, d = .44) to Very Large (VWMI, d = 1.20) with a mean effect size of 0.75 (Large) at the index score level.

At the subtest level, U-SIM participants also performed quite poorly on the verbal subtests. Logical Memory Immediate (LM1) scaled scores ranged from 1 to 12 with 85.4% of participants demonstrating impaired performance. On LM2, scores ranged from 1 to 13 with 90.2% of scores falling in the impaired range. For this subtest, 14.6% and 7.3% of U-SIM participants scored in the Average range on immediate and delayed recall portions respectively. In comparison to Good Effort TBI participants, U-SIM participants performed significantly worse on both immediate recall (F (1, 80) = 4.496, p = .037) and delayed recall (F (1, 80) = 5.266, p = .024). The pooled SD effect sizes were classified as medium effects for both immediate (d = .47) and delayed (d = .48) recall. Logical Memory recognition (LM-Rec) performance was also poor; raw scores ranged from 10 to 29 with 51.2% of the sample achieving 65% accuracy.

Performance on VPA was similar with 68.3% of scores in the impaired range for VPA1; on VPA2, 58.5% of scores were classified as impaired. There were no significant differences between U-SIM and good effort TBI participants on VPA performance for either the immediate or delayed recall. On VPA recognition (VPA-Rec), the total number of correct responses for the U-SIM sample was fairly evenly distributed with raw scores ranging from 11 to 40 and over 80% of the sample demonstrating greater than 65% accuracy. More refined analysis of VPA-Rec performance suggests quite poor performance. Although the average total hit rate was high (M = 10.1, SD = 3.7), discriminability ranged from .26 to 1.00 with a mean of .76 (SD = .18). More than half of U-SIM participants correctly identified all of the easy items (56.1%), but only 56.1% correctly identified 8 out of 10 hard items, and the number of false-positive errors was high (M = 5.0, SD = 4.8). In comparison to Good Effort TBI participants, U-SIM participants demonstrated significantly lower discriminability scores (F (1, 79) = 13.45, p < .001, η^2 = .15).

On the visually-based immediate and delayed recall subtests, U-SIM participants again demonstrated poor performance. For Designs 1 (DES1), scores ranged from 1 to 11 with 65.9% of scores in the impaired range. Designs 2 (DES2) performance ranged from 1 to 13 and found 63.4% of scores in the impaired range. In comparison to good effort TBI participants, U-SIM performance was no different on the immediate recall portion (although a trend was observed; (F (1, 80 = 3.46, p = .067, η^2 = .04), but was significantly worse on the delay portion (F (1, 80) = 10.09, p = .002, η^2 = .11). Visual reproduction Immediate Standard Scores (SS) ranged from 1 to 12 with a surprising 31.7% of participants achieving a SS of 1 on this subtest, suggesting that this subtest in particular was a common target of dissimulation strategy. A total of 75.6% of scores were in the impaired range. The same pattern was not observed for Visual Reproduction-2 (VR2) with scores ranging from 1 to 15. Only 7.3% achieved a SS of 1 and 80.5% scored in the impaired range. Unsuccessful SIM performance was significantly lower in comparison to Good Effort TBI for both immediate (F (1, 80) = 15.95, $p < .001$, $\eta^2 = .17$) and delay conditions (F (1, 80) = 13.54, $p < .001$, $\eta^2 = .15$). On the recognition portion of the visual memory subtests, 51.2% of U-SIM participants achieved 50% accuracy on Designs-Recognition (DES-Rec) and on Visual Reproduction-Recognition (VR-Rec), 75.6% of the sample correctly identified only three or fewer of the designs. Simulator recognition performance was significantly lower for both Designs (F (1, 79) = 6.56, $p = .012$, $\eta^2 = .08$) and VR (F (1, 80) = 67.69, $p < .001$, $\eta^2 = .49$) in comparison to good effort TBI.

The two additional working memory subtests (Spatial Addition and Spatial Span) also revealed a consistent pattern of impaired performance with SS's ranging from 1 to 10 for both Spatial Addition (SA) and Spatial Span (SSp). On the SA subtest, 17.1% of participants achieved a SS of 1 with 85.4% of scores in the Impaired range. On SSp, the modal SS was 6 with 82.9% of scores in the impaired range. Simulators again performed significantly worse on these two subtests (SA, F (1, 80) = 19.77, $p < .001$, $\eta^2 = .20$; SSp, F (1, 80) = 21.47, $p < .001$, $\eta^2 = .21$) in comparison to good effort TBI.

On the BCSE, U-SIM participant scores ranged from 2 to 56 with 61.0% of participants scoring in the Very Low range. Only 2.4% U-SIM participant scored in the Average range, 7.3% scored in the Low Average range, 12.2% in the Borderline range, and 17.1% in the Low range. BCSE scores for the U-SIM group were significantly lower than the good effort TBI group (Mann-Whitney U = 211.5, $p < .001$) with an effect size of 1.55 (huge).

Averaging across the effect sizes (Cohen's *d*) for the individual subtests yielded a mean of 0.61 (medium effect). The mean effect size for the immediate recall verbal subtests, including the substituted VPA scores was 0.32 (small) and for verbal delay was 0.42 (small). Visual subtests demonstrated larger effects with a mean effect size of 0.71 (medium) for immediate visual subtests and 0.77 (medium) for delay subtests. The working memory subtests, however, demonstrated the largest effect sizes (SA = 1.03 and SSp = 1.04). Effect sizes for most of the recognition subtests were quite large with a mean of .83 (large) and range from .32 (LM-Rec) to 1.83 (VR-Rec).

In comparison to the Suspect Effort TBI group, there were no significant differences in the number of impaired index scores using either method, and the only index score that demonstrated a significant difference was the standard administration of the AMI (F (1, 56) = 5.06, $p = .028$, $\eta^2 = .08$). More specifically, suspect effort TBI participants performed significantly worse on this index than the unsuccessful simulator participants. All other pairwise index score comparisons between Suspect Effort TBI and U-SIM participants were not significant.

At the subtest level, the only significant difference observed was for VPA. More specifically, suspect effort TBI participants scored significantly worse on VPA1 (F (1, 57) = 17.97, $p < .001$, $\eta^2 = .24$) and on VPA2 (F (1, 56) = 7.52, $p = .008$, $\eta^2 = .12$), which contributed to the significantly lower AMI. Across all other subtests, no significant differences were observed between these two groups. In terms of recognition performance, no significant differences were observed for any of the 4 recognition portions.

Section 5.2.1.4 Successful SIM

The Successful Simulator (S-SIM) group is composed of healthy adults who reported feigning memory impairment on the debriefing questionnaire, but who passed at least five of the six effort measures. As can be seen in Table 3, all of the Index scores and subtest SS means were within 1 SD of the mean. In this sample, the number of impaired index scores using the standard calculation ranged from 0 to 5 with most participants generating no impaired index scores; 50.0% of this sample demonstrated impairment on two or more index scores. Using the substitution method, the most participants in this group did not demonstrate impairment on any index score; 44.4% of this sample showed impairment on two or more of the index scores using CVLT-II indices. On any of the individual subtests, no more than 50% of the S-SIM sample demonstrated impairment.

Section 5.2.2 – Effort Measures

Effort test performance by group is presented in Table 6 (TOMM, RDS, and CVLT-II indices) and Table 7 (Green Measures). Of the four groups, the U-SIM group had the highest failure rate on four of the six effort measures. The exceptions were the CVLT-II three-variable model and the NV-MSVT on which the Suspect Effort TBI group had a slightly higher failure rate. The S-SIM group had the lowest failure rate on all effort measures except for the NV-MSVT. In the Good Effort TBI group, 70.7% of participants failed one measure with most failures on the CVLT-II three-variable model; 29.3% passed all six effort measures. In the suspect effort TBI group, 47.4% of participants failed two SVT's, 36.8% failed three, and 15.8% failed four of the six SVT's; the highest failure rate for this group was observed on the NV-MSVT.

The majority (66.7%) of participants in the S-SIM group passed all six of the effort measures with only 33.3% failing one SVT; the NV-MSVT was the most failed measure among S-SIM participants. The number of failed SVT's for the U-SIM group ranged from 2 (19.5%) to all 6 (19.5%) with most U-SIM participants failing four of the six (26.8%). The highest failure rate among U-SIM participants was observed on the NV-MSVT.

Section 5.3 – Primary Prediction Equations

Hypothesis 1: It was predicted that individual indices or a combination of indices obtained from administration of the WMS-IV would be capable of reliable differentiation between persons with verified TBI and neurologically-normal participants dissimulating TBI.

Section 5.3.1 – Advanced Clinical Solutions Embedded Measures

The first set of predictors used were those included in the Advanced Clinical Solutions (ACS) package as embedded effort indicators. For the following ACS analyses only, which includes raw score and base rate models, the RDS and WCT were not used to classify subjects as putting forth good or poor effort to avoid contamination of predictor with criterion, because these two measures are included in the ACS. Following classification of effort using the previously described methods without inclusion of the RDS, the resulting Good Effort TBI sample (n = 45) was predominantly African American (76%) men (91%) with a mean age of 43.4 years (SD = 12.1) and a mean education of 11.8 years (SD = 2.3). The Unsuccessful SIM group (n = 39) was also predominantly African American (74%) men (82%) with a mean age of 42.4 years (*SD* = 11.5) and mean education of 12.8 years (*SD* = 2.0). These two groups were initially equivalent on age, gender, and ethnicity, but with this reclassification, the SIM group demonstrated a significantly higher mean level of education (F (1, 82) = 4.49, p = .037). In regard to test performance, the SIM group performed significantly worse than the TBI group on four of the five ACS variables; the exception was LM-Rec which did not significantly differ between these two groups (F (1, 82) = 0.86, p = 36). Test performance data for this analysis are presented in Table 8 along with effect sizes (Cohen's *d*) calculated using a pooled standard deviation.

An initial model was fitted using each of the five individual embedded effort measures identified by Pearson (e.g., WAIS-IV Reliable Digit Span, Word Choice Test, Logical Memory recognition, Visual Reproduction recognition, and Verbal Paired Associates recognition) as predictors with group membership (e.g., Good Effort TBI vs. Unsuccessful SIM) as the outcome of interest. Education was also entered to account for the statistical difference observed between groups using this particular classification scheme. A test of this model against a constant only model was statistically reliable (χ^2 = 72.38, p < .001, Nagelkerke R^2 = .77) with an AUC of .95 (95% CI = .90 - 1.00), indicating that these five measures together with education performed extremely well at discriminating between groups in the present sample. The overall hit rate for this model was 90.5% with 89.7% sensitivity to poor effort and 91.1% specificity. Although the overall model was statistically reliable, only four of the six individual predictors were significant. More specifically, education and LM-Rec did not contribute significantly to the overall model and participant classification, thus reinforcing the utility of the remaining four variables. The ROC curve for this model is presented in Figure 2 and model summary statistics for the Full ACS model with education are presented in Table 9.

Note: The remaining analyses use the primary participant classification described in the methods section (i.e., use of all effort measures except for the Word Choice test). As a result, the model fit statistics used in the present study (e.g., Bayesian Information Criterion) for the previous model cannot be compared to those for the subsequent models because the Pearson ACS model was fit using a slightly different sample. Rather, qualitative comparisons can be made using AUC values.

Section 5.3.2 – Word Choice Test

To examine the efficiency of the Word Choice test (WCT) separately, a model was fit using only the WCT raw score. A test of this model against an empty model was statistically reliable (χ^2 = 32.89, p < .001, Nagelkerke R^2 = .41) with an AUC of .83 (95% CI = .74 - .92). The odds ratio for the Word Choice test was 0.82 (p < .001; 95% C.I. = 0.74 - 0.90). When interpreting odds ratios, a value of 1.00 is indicative of little to no predictive ability and the number translates to a percent change in probability of group membership. As values depart from one – in either direction – the greater the degree of influence a particular variable has on prediction of the outcome. In this case, an odd's ratio of .817 indicates that for every unit increase in the score on the WCT, the probability of being classified as a simulator decreases by 18.3%. The ROC curve for this model is presented in Figure 3.

Based on predicted probabilities using a criterion of .50 (which translates to a cut off score of 43), the WCT correctly classified 78.0% of the sample and demonstrated an overall sensitivity to simulation of 73.2% with 82.9% specificity. Examining the coordinates of the ROC curve revealed that the diagnostic efficiency can be improved by using a lower cutoff score of 37 that yielded 56.1% sensitivity to poor effort with 95.1% specificity for the present sample. Given the risk associated with false-positives (i.e., falsely classifying an individual as malingering), the lower cutoff generating higher specificity is preferred, even if it comes at the expense of sensitivity.

The Bayesian Information Criterion (BIC) was also calculated in order to generate an overall estimate of model fit for use in comparing to subsequent models fit to the same sample that are not nested. The BIC is a model fit statistic used to compare prediction models (Hardin & Hilbe, 2007) based on the deviance statistic (D) of a specified model (M_K), sample size (n), and degree's of freedom (df). It is calculated as:

$$BIC = D\,(M_K) - (df)\ln(n)$$

For the previous model using only the raw score from the WCT, the calculated BIC was -271.75. Although a BIC value in isolation is relatively uninformative and there is no statistical test associated with it, it is very useful when it can be compared to other BIC values calculated from models fit to the same data. According to interpretive guidelines, the more negative value is associated with the preferred model, and the difference between BIC values can be used as an indication of the degree of preference. Differences between 0 and 2 indicated a “Weak” preference for the model with the more negative value, 2 to 6 yield a “Positive” preference, 6 to 10 suggests

"Strong" preference, and differences greater than 10 provides "Very Strong" evidence (Raftery, 1996).

Section 5.3.3 – Wechsler Memory Scale, Fourth Ed. Brief Cognitive Status Exam

Included in the WMS-IV as an optional subtest is the Brief Cognitive Status Exam (BCSE). The BCSE is a measure intended to provide an overview of cognitive functioning including orientation, language, visual-spatial, memory, and inhibitory abilities and is a replacement for the Information and Orientation subtest from the WMS-III. Administration generates a total raw score that along with age and education, is used to classify test takers as Average, Low Average, Borderline, Low, or Very Low. Although the classification grouping is too coarse a measure to use as a predictor, the total raw score represents a viable option as a measure of symptom validity.

Based on observations made during data collection as well as the rather large effect size found, the BCSE was selected as a candidate for use as a variable for the detection of suboptimal effort. A model was fit that again used the Unsuccessful SIM and Good Effort TBI as the groups of interest and primary outcome variable and the total raw score for the BCSE was used as the predictor. A test of this model against a constant only model was significant ($\chi^2 = 40.24$, $p < .001$, Nagelkerke $R^2 = .52$) with an AUC of .87 (95% CI = .80 - .95). In this model, the BCSE was a significant predictor with an odds ratio of 0.85 (95% C.I. = 0.79 – 0.92). The BIC value associated with this model was -279.10. Examining the coordinates of the ROC curve finds that using a cut off score of 34 that yielded 52.5% sensitivity to simulation and 93.9% specificity maximized diagnostic efficiency. The ROC curve for this model is presented in Figure 4.

Section 5.3.4 – Wechsler Memory Scale, Fourth Ed. Visual Reproduction Recognition

Although VR-Rec is included in the ACS package, it was evaluated separately given the very large effect size reported in earlier analyses, suggesting that it may be a viable option independently. Fitting a model with VR-Rec as a predictor and group membership as the outcome of interest was significant (χ^2 = 45.58, *p* < .001, Nagelkerke R^2 = .57) with an AUC of .88 (95% CI = .80 - .96). The odd's ratio for VR-Rec was 0.37 (*p* < .001; 95% C.I. = 0.25 - 0.55) and the BIC value for this model was -284.44. With an odd's ratio so discrepant from 1.00, it is clear that VR-Rec had a very large effect on classification. Examining the coordinates of the curve suggest that a cutoff score of 3 was the most appropriate. At this level, sensitivity to simulation was 61.0% and specificity was 96.4%; the ROC curve is presented in Figure 5.

Section 5.3.5 – Wechsler Memory Scale, Fourth Ed. Verbal Paired Associates Recognition

As with VR-Rec, Verbal Paired Associates recognition (VPA-Rec) is also included in the ACS package. However, the ACS uses only the total raw score and makes no consideration of the process measures available (e.g., easy hits, hard hits, false positives, discriminability). Given the availability of these measures and a history of their effective use in effort research with the CVLT and CVLT-II (e.g., Millis, Putnam, Adams, & Ricker, 1995; Wolfe et al., 2010), prediction models were fit using these variables.

The first model fit using these indices contained raw scores for total hits and total false-positives. A test of this model against a constant only model was significant (χ^2 = 15.33, *p* < .001, Nagelkerke R^2 = .23), but the only predictor to contribute significantly to the model was the total number of hits (Odds Ratio = 0.71, p = .007; 95% C.I. = 0.56 - 0.91); the total number of false positives was not significant. This model was deconstructed and refitted by breaking down the total number of hits into the number of easy and hard hits and using those values as predictors instead. A test of this model against an empty model was also significant (χ^2 = 11.88, p = .003, Nagelkerke R^2 = .18), but the only predictor to contribute to the overall model was the total number of hard hits (Odds Ratio = 0.73, p = .012; 95% C.I. = 0.57 - 0.93); the number of easy hits was not significant. The model was again refitted, but this time only the number of hard hits was entered as a predictor. This model was significant (χ^2 = 11.46, p = .001, Nagelkerke R^2 = .17) and the number of hard hits was a reliable predictor on its own (Odds Ratio = 0.71, p = .004; 95% C.I. = 0.56 - 0.90). The BIC for the model containing only the number of hard hits was -250.32 and the AUC was .70 (95% C.I. = .59 - .82).

In the continued search of VPA-recognition process variables to be used as predictors of group membership, a model was fit using VPA-Rec discriminability. This model was also significant (χ^2 = 13.71, p < .001, Nagelkerke R^2 = .21) as was the only predictor (Odds Ratio = 0.002, p = .002; 95% C.I. = 0.000 - 0.100). The calculated BIC value for the discriminability only model was -248.59. As with the number of hard-hits, the AUC for this model was disappointingly low at .72 (95% C.I. = .61 - .83). Examining the coordinates of this curve reveal that a discriminability cut-off of .80 achieves sensitivity to poor effort of 54% with 90% specificity. In comparison to the hard-hits model presented previously, the difference in BIC values indicates a “Weak” degree of preference for the model using the number of hard hits; however, the larger AUC for discriminability suggests better overall diagnostic efficiency. The ROC Curves for both of the VPA recognition variables are presented in Figure 6 simultaneously.

Section 5.3.6 – Wechsler Memory Scale, 4th Ed., Spatial Span Subtest

The Spatial Span (SSp) subtest is a measure that is a visual analog to the Digit Span subtest from the WAIS-IV. Given the similarity, it was hypothesized that a measure of effort could be calculated using the longest span correctly recalled. For each participant, the number of items in the last correctly-recalled item was recorded and entered into a prediction model. Although the WAIS-IV uses a “reliable” span, the longest correctly recalled span from the SSp was chosen because there are up to five trials per span, and given the low level of performance, using a reliable span would have caused a severely restricted range. The average longest span for the Good Effort TBI group was 3.2 (*SD* = 0.5) and for the U-SIM group was 2.5 (*SD* = 0.9); this difference was significant (F (1, 80) = 22.98; p < .001, η^2 = .22).

When entered into a logistic regression, a test of this model was statistically reliable (χ^2 = 21.59, p < .001, Nagelkerke R^2 = .31) with an AUC of .750 (95% C.I. = .644 - .855). The odds ratio for the longest span was 0.21 (p < .001; 95% C.I. = 0.09 - 0.48); the BIC value associated with this model was -260.45, indicating a high degree of fit in comparison to other indices towards this hypothesis. Analysis of the coordinates of the ROC curve identified that a cutoff score of 2 yielded a 30.4% sensitivity, which is unacceptably low, but a near perfect specificity of 97.5%. When this value was raised to a cut score of 3, sensitivity increased to 68.3%, but specificity dropped considerably to 63.4%. Given the risk of false positives in relation to false negatives, a more conservative cutoff score of 2 would be advisable, given the very high specificity. With such a low sensitivity, it would be essential to use this measure only in conjunction with other, well-validated measures.

Section 5.3.7 – The Advanced Clinical Solutions, Revised

In light of the previous set of findings, the set of variables included in the ACS was revisited and revised. Based on the classification accuracy of the variables explored, a model was fit that contained those variables identified as robust as well as those hypothesized to be among the most commonly employed in clinical practice. The outcome of interest was again group membership and the variables entered as predictors were the BCSE total score, Word Choice Test raw score, VR-Rec total raw, and VPA-Rec discriminability. These variables were selected based on their performance individually, and the likelihood that they would be among the most commonly used measures by clinicians. All variables were entered simultaneously and a test of this model against a constant only model was significant (χ^2 = 67.93, p < .001, Nagelkerke R^2 = .76) with a BIC of -289.63; all four predictors were significant. The AUC for this model was .95 (95% C.I. = .91 - .99) indicating an "outstanding" degree of diagnostic efficiency and overall classification accuracy (Hosmer & Lemeshow, 2000).

Inspection of the odds ratios, however, suggested the possibility for the presence of multicollinearity. More specifically, VPA-Rec demonstrated an unusually large odds ratio and an extremely wide confidence interval that was disproportionate to the scale of measurement and the remaining odds ratios. Collinearity diagnostics were run and revealed the presence of significant overlap between the Word Choice Test and VPA-Rec discriminability. Thus, a decision was made to remove the WCT from the model because it was previously evaluated independently, and the model was refitted using the BCSE total raw score, VR-Rec and VPA-Rec discriminability. A test of this model against and empty model was significant (χ^2 = 58.75, p < .001, Nagelkerke R^2 = .69) with an AUC of .93 (95% C.I. = .87 - .99); two of the three predictors were significant. The BIC value for this model was -284.84. Model summary statistics including regression coefficients and odds ratios for the individual predictors are presented in Table 9, and the ROC curve is presented in Figure 7.

Section 5.4 – Secondary Prediction Equations

The models investigated in Hypothesis 1 were revisited to determine their abilities to address Hypothesis 2. Then additional exploration determined whether other viable models existed within the present data. Due to the small size of the successful simulator group (n = 18), models with more than two predictors should be interpreted with caution and are not advised for clinical use until replicated with a larger sample.

Hypothesis 2: It was predicted that individual indices or a combination of indices obtained from administration of the WMS-IV would be capable of reliably differentiating between successful TBI dissimulators (S-SIM) and unsuccessful TBI dissimulators (U-SIM)

Section 5.4.1 – Advanced Clinical Solutions Embedded Measures

The first model fit using all five raw scores from the ACS was statistically reliable in comparison to an empty model (χ^2 = 41.39, p < .001, Nagelkerke R^2 = .71); however, none of the individual predictors themselves were significant, suggesting the presence of multicollinearity. Examining collinearity diagnostics revealed a substantial amount of overlap between the Word Choice test and VPA recognition, which is consistent with the previous model that found overlap between VPA-Rec discriminability and the WCT. As in the previous analysis, given that the WCT would be evaluated separately, a decision was made to refit the model without inclusion of the WCT and retain the embedded nature of the ACS.

Refitting the model with only the three recognition variables and the RDS was statistically reliable (χ^2 = 37.34, p < .001, Nagelkerke R^2 = .66) with an AUC of .94 (95% C.I. = .88 - .99) indicating a high degree of classification accuracy; the BIC for this model was -184.95. This reduced version of the ACS model correctly classified 84.7% of participants with 72.2% sensitivity to successful simulation and 90.2% specificity. Out of the four variables entered, only the RDS and VR-Rec were significant predictors; LM-Rec and VPA-Rec were not significant. This finding is consistent with observations made in the analyses addressing Hypothesis 1. However, it is important to consider the available power in the present sample to effectively fit a model with so many predictors; it is presented here to provide preliminary evidence of its utility. As a result, with a larger sample the previous model may in fact be a better fit. Summary statistics are presented for the ACS 4-Variable model in Table 10; however, the use of regression coefficients to generate predictions from this model is ill advised. The ROC curve for this model is presented in Figure 8.

Section 5.4.2 – Word Choice Test

The Word Choice test alone was also a reliable predictor between successful and unsuccessful simulators (χ^2 = 28.97, p < .001, Nagelkerke R^2 = .55), with an odds ratio of 0.72 (95% C.I. = 0.59 - 0.89). The BIC value for this model was -188.81, and the AUC for the Word Choice test as a predictor was .90 (95% C.I. = .81 - .98), demonstrating a high degree of diagnostic efficiency. Evaluating the coordinates of the curve find that a modest balance between sensitivity and specificity was achieved using a cutoff score of 37, which yielded a sensitivity to successful simulation of 94.4% and a specificity of 56.1% Interestingly, cutoff scores between 36 and 46 all yielded the same sensitivity, but specificity varied dramatically. At the upper end, using a cutoff of 46, specificity to successful simulation was maximized at 78.1% while retaining the 94.4% sensitivity. The high degree of sensitivity suggests that this measure is very good at detecting cases of poor effort; however, it is inefficient at differentiating between those who simulated well and those who did not. The ROC curve for the WCT only model is presented in Figure 9.

Section 5.4.3 – Wechsler Memory Scale, Fourth Ed. Brief Cognitive Status Exam

Using the BCSE to differentiate between successful and unsuccessful simulators was statistically reliably in comparison to a constant only model (χ^2 = 14.97, p < .001, Nagelkerke R^2 = .32) with an AUC of .81 (Figure 10; 95% C.I. = .68 - .94). The odds ratio for the BCSE in this model was 0.90 (p = .002; 95% C.I. = 0.85 - 0.97) and the overall hit rate was 83.1%. Specifying a cutoff score of 45 yielded 63.9% sensitivity to successful simulation and 83.0% specificity. Raising the cutoff to 47 lowered sensitivity to 58.4%, but raised specificity to 92.7%. This cutoff score is quite high; however, considering the performance of the TBI group in the present sample, as well as the ease of this measure, a score of 45 (out of 58) is actually quite poor and would be classified as Borderline or Low Average for most participants, except for those with an 8th grade education or less. The BIC value for this model was -174.81, which suggests that although the BCSE is a statistically reliable predictor, it is not as good a fit as the WCT is between the two simulator samples.

Section 5.4.4 – Wechsler Memory Scale, Fourth Ed. Visual Reproduction Recognition

Visual Reproduction recognition was again a very reliable predictor of group membership (χ^2 = 24.12, p < .001, Nagelkerke R^2 = .47) with an odds ratio of 0.44 (p < .001; 95% C.I. = 0.29 - 0.67) and an AUC of .86 (95% C.I. = .75 - .97). Using a cutoff score of 5 out of 7 was necessary to achieve sensitivity to successful simulation of 77.8% and specificity of 84.2%. Increasing the cutoff to 6 lowered sensitivity to 47.3% but raised specificity to 91.5%. The BIC value for this variable was -183.96, which thus far was the second-best fit of the present data of the single-variable prediction models.

Section 5.4.5 – Wechsler Memory Scale, Fourth Ed. Verbal Paired Associates Recognition

Process variables from VPA-Rec were again entered into prediction models with simulator status as the outcome of interest. Given the findings from section 1, two separate models were fit; one using VPA-Rec Hard Hits and one using discriminability. The first model using the number of hard hits was statistically reliable (χ^2 = 15.60, p < .001, Nagelkerke R^2 = .34) with an odds ratio of 0.45 (p = .009; 95% C.I. = 0.25 - 0.82) and an AUC of .79 (95% C.I. = .68 - .91). The BIC for this model was -172.82. A very high cutoff score was necessary to achieve an adequate balance between sensitivity and specificity. More specifically, using a cutoff of 9 (out of 10) resulted in 73.6% sensitivity to successful simulation and 70.8% specificity. Using a lower cutoff score of 6 resulted in 100% sensitivity; however, specificity was low at 31.7%. A cutoff of 7 yielded a sensitivity of 97.1% and specificity of 39.0%. Thus, as with the WCT, a low number of hard hits on VPA recognition is very sensitive to simulation, but not very specific to the level of sophistication.

Using VPA-Rec discriminability was also a reliable predictor (χ^2 = 23.02, p < .001, Nagelkerke R^2 = .47) with an odds ratio of 0.00 (95% C.I. = 0.00 - 0.01) and an AUC of .86 (95% C.I. = .77 - .96); the BIC value for this model was -180.23. Using a cutoff score of .95 – near perfect discrimination – resulted in 52.9% sensitivity to successful simulation and 87.8% specificity. Using a discriminability score of .74 as a cutoff, perfect sensitivity was achieved, but specificity dropped to 39.0%. Although each of these variables reliably differentiated between groups, very stringent cutoff scores were necessary to achieve acceptable diagnostic efficiency suggesting that the S-SIM group performed well on these measures in comparison to the U-SIM group, at levels which are approaching an intact level of performance. Thus, as with the other variables identified in the present section, the VPA-Rec Discriminability score was effective at capturing simulators, but not very efficient at differentiating between those who simulated well, and those who do not.

Section 5.4.6 – Longest Spatial Span

Given the performance of the longest spatial span in section 1, its ability to differentiate among simulator samples was also evaluated. The overall model was significant as compared to an empty model (χ^2 = 18.43, p < .001, Nagelkerke R^2 = .38) with an AUC of .81 (95% C.I. = .68 - .93). The only predictor was significant (p = .001) with an odds ratio of 0.20 (95% C.I. = 0.08 - 0.52). The BIC for this model was -178.27, and in comparison to the previously fit models, does not represent the best fit of the available data. The diagnostic efficiency of the longest span was not much better in the present section than in section 1. Using a cutoff score of 5 achieved perfect specificity;

however, sensitivity to successful simulation was virtually zero. Using a cut off of 3, however, yielded 75.0% sensitivity and 68.3% specificity.

Section 5.4.7 – The ACS-Revised Model

In this section, the revised ACS model identified in section 1 was evaluated for its ability to differentiate between successful and unsuccessful simulators The overall model was significant as compared to an empty model (χ^2 = 30.70, p < .001, Nagelkerke R^2 = .59) with an AUC of .91 (95% C.I. = .82 – 1.00) and of the three predictors entered, VPA-Rec discriminability was significant (p = .046; Odds Ratio = 0.00, 95% C.I. = 0.00 - 0.81) and VR-Rec was also significant (p = .021; Odds Ratio = 0.56, 95% C.I. = 0.34 - 0.92); the BCSE total score was not significant. The BIC for this model was -179.79, and in comparison to the previously fit models, it did not represent the best fit of the available data. It is important to note, however, that this particular model approaches the limits of available power with three predictor variables entered. In the presence of a larger successful simulator sample, this model may be a better fit. Model summary statistics are not reported for this model due to its instability, but the fit statistics are presented in Table 11 for comparison to the same model used in section 1. The ROC curve, however, is plotted and can be found in Figure 11.

Section 5.4.8 – The 2-Varible ACS Model

In order to conform to the constraints of the S-SIM sample size, the previously fit variables were explored, and based on BIC values and odds ratios when entered individually, the two best fitting indices were simultaneously entered; the resultant model included the WCT and VR-Rec as predictors. This two-variable model was statistically reliable (χ^2 = 35.86, p < .001, Nagelkerke R^2 = .64) with an AUC of .94 (95% C.I. = .87 -

.99). The BIC value for this model was -191.62, the largest observed thus far for models addressing Hypothesis 2. Both predictors were statistically significant with odds ratios of less than 1. Model summary statistics are presented in Table 10, and the resulting ROC curve is presented in Figure 12. Comparing the overall performance of this model to the remaining models fit towards addressing this hypothesis, this model is the best fit. Interestingly, it demonstrated nearly identical diagnostic efficiency as the four-variable model in this section, and in comparison to the Revised ACS model fit in section 1, its performance was marginally lower based on comparison of ROC curves.

Section 5.5 – Additional Prediction Equations

Although not specified as an a priori model for section 1, the two-variable model was also fit using good effort TBI and U-SIM classification as the outcome of interest as a basis for direct comparison. A test of this model against a constant only model was significant (χ^2 = 51.91, p < .001, Nagelkerke R^2 = .63) with an AUC of .91 (95% C.I. = .85 - .97). The BIC value for this model was -286.36. In comparison to the other models fit for Hypothesis 1, this model did not perform as well, although it was the second-best fitting model tested. Comparing the performance of the two-variable model between Hypothesis 1 and Hypothesis 2, based on AUC values, it appears that this model is more effective at differentiating between cases of successful and unsuccessful simulation than between cases of unsuccessful simulation and good effort TBI.

Even though not part of the original project proposal, a third set of models was fit towards addressing a third hypothesis which is of substantial clinical interest; specifically, whether or not the previously identified models are capable of differentiating between successful simulators and good effort TBI. Given the limited sample size of the successful simulator group, and to avoid fitting underpowered models, only the individual predictors were summarized as well as models that contained two predictors or fewer. Of all the variables that reliably differentiated between groups in the previous two sections, the *only* variable that reliably differentiated between successful simulators and good effort TBI was VPA recognition discriminability (χ^2 = 7.99, p = .005, Nagelkerke R^2 = .19) with an AUC of .74 (95% C.I. = .60 - .88), which did not perform particularly efficiently and is barely considered acceptable clinical use (Hosmer & Lemeshow, 2000). Several of the other predictors approached significance (e.g., Word Choice test, VPA Recognition Hard Hits, and the longest Spatial Span), which suggested given a larger sample they may be viable predictors.

CHAPTER 6

DISCUSSION

Overall, the findings support the hypothesis that the fourth edition of the Wechsler Memory Scale (WMS-IV) can be a reliable and valid tool to identify suboptimal effort and feigned memory impairment. Indices from the WMS-IV reliably differentiated between individuals coached to feign cognitive impairment that were clearly withholding effort (unsuccessful simulators) and traumatic brain injury (TBI) survivors putting forth good effort. When patterns of performance were evaluated across several of these indices, the predictive abilities were improved. Similarly, individual indices as well as patterns of performance across multiple indices from the WMS-IV reliably differentiated between coached individuals who did so successfully (i.e., avoided detection) and coached individuals who were less sophisticated in their simulation strategies (i.e., were identified as feigning). Furthermore, indices were identified that offer promise towards differentiating between TBI survivors and coached individuals who avoided detection, which is of considerable clinical interest. Thus, based on these findings, it can be concluded with relative certainty that the WMS-IV contains indices obtained from routine administration that can be used for the detection of negative response bias and suboptimal effort.

Section 6.1 – Hypothesis 1 (Good Effort TBI vs. Unsuccessful Simulator)

The present study not only confirmed that the WMS-IV can detect cases of symptom exaggeration, it offers specific measures that will allow clinicians to do so. A primary focus of the present study was to determine the ability of the WMS-IV to differentiate between cases of individuals with bona fide TBI from healthy adults who were coached to feign cognitive impairment associated with TBI. Being able to do so in a clinical setting is of great interest to clinicians and researchers alike, especially with the relatively high base rate of malingering (Larrabee et al., 2009; Mittenberg et al., 2002). Because memory is a common target of feigning strategies (Suhr & Barrash, 2007), the WMS-IV is therefore highly susceptible to symptom exaggeration. Given the very large number of potential scores and values that can be used to classify individuals, only raw scores and scaled scores were used with few exceptions which were made based on previous research (e.g., discriminability, longest spatial span). Although these methods are by no means intended to be exhaustive, they are an excellent point from which further research can be implemented and based.

One of the best fitting models was the five-variable model included as part of that Advanced Clinical Solutions (ACS) package that evaluated performance across multiple measures within the WMS-IV. After the release of the WMS-IV, Pearson Education released the ACS package that contained two stand-alone symptom validity tests and identified three WMS-IV recognition variables as embedded measures of effort. The two stand-alone measures were the Reliable Digit Span (RDS) developed by Greiffenstein, Gola, & Baker (1995) from the Wechsler Adult Intelligence Scale, 4th Ed. (WAIS-IV), and the Warrington Recognition Memory test (Warrington, 1984) analog, the Word Choice test (WCT). The RDS is not a new measure, and it is not surprising that it was included in this package considering the long-standing history of use as an embedded measure of symptom validity (e.g., Axelrod et al., 2006; Larrabee, 2003; Mathias, Greve, Bianchini, Houston, & Crouch, 2002; Nelson, Boone, Dueck, Wagener, Lu, & Grills, 2003). Pearson’s use of recognition measures was also consistent with other embedded symptom validity indicators. Typically, recognition paradigms create a forced-choice response option that allows for calculation of chance-performance levels (Greve, Binder, & Bianchini, 2009; Hiscock & Hiscock, 1989), and moreover, recognition memory measures have demonstrated previous efficacy in this area. Their production of the WCT is also consistent with previous methods of assessing symptom validity. Similar measures have been used for some time (e.g., Warrington Recognition Memory Test; Warrington, 1984) and the WCT performed much like its predecessor (Millis, 1994).

When all five of the measures included in the ACS method were used in conjunction with education, the group of measures was highly successful at accurately classifying individuals as an unsuccessful simulator or a true TBI survivor putting forth good effort, offering strong evidence in support of Hypothesis 1. The error rate for this model was less than 10%, and most misclassification errors were false negatives (i.e., missed cases of malingering) as opposed to false positives (i.e., bona fide TBI misclassified as a faker). If used in a clinical setting, false-positive errors carry greater consequence and thus, this model balances types of errors well, and in the preferred direction. Of note, education had very little impact on classification. In fact, education showed weak relation to any of the WMS-IV indices that would be used to identify suboptimal effort. Of the five indices included in the ACS classification model, four met rigorous criteria for contributing to accurate classification of the groups, with the exception of Logical Memory recognition.

For Logical Memory recognition, participants classified as unsuccessful simulators generally performed comparably to TBI survivors putting forth good effort, and based on effect sizes, Logical Memory recognition was least affected by coaching of the five ACS measures. Overall, the unsuccessful simulators were detected as such because they overestimated the impairment that persons with TBI would show on the other ACS measures. However, they did well in approximating the amount of impairment on Logical Memory recognition; in other words they did poorly on the task, sufficient to appear brain injured, but not so poorly as to appear disproportionately impaired. What remains unclear is whether or not the TBI group performed particularly poorly on this measure, or alternatively if the unsuccessful simulator group selectively withheld their strategy on this task. Another possibility is that scoring rubric for Logical Memory recognition (total hits) is an insensitive method to detect the simulators' strategy, and further refinement of this score is necessary as was done with Verbal Paired Associates Recognition. Similar approaches have been taken with the previous edition of the Wechsler Memory Scale and the Rarely Missed Index (Killgore & DellaPietra, 2000), though such efforts have met mixed review (Axelrod, Barlow, & Paradee, 2010; Bortnik, Boone, Marion, Amano, Ziegler, Victor, & Zeller, 2010).

An ideal method to address this issue would be to convert the raw scores to a common metric using the means and standard deviations from the normative sample, but no such data are available. Alternatively, base rates of performance could be examined for each individual using the groupings assigned by Pearson; this approach, however, is a coarse metric. Nonetheless, it is noteworthy that the base rate of the average performance for the TBI group was at the 17th to 25th percentile on Logical Memory recognition, which is considered marginally poor. The SIM group's base rate classification was considerably lower, falling at the 3rd to 9th percentile, thus arguing against the hypothesis that simulators changed their response style on this task. It is important to note that performance on this measure is highly negatively skewed, such that even a single point difference can have a dramatic effect on base rate performance, and there were no differences in performance base on raw scores. In sum, it appears that the Logical Memory recognition score was not more easily or differently feigned, rather it was a relatively challenging task for persons with TBI which narrowed the typical gap from simulators, who usually grossly overestimate the level of impairment.

Each of the other four indices in the ACS method (Reliable Digit Span, Word Choice, Verbal Paired Associates recognition, and Visual Reproduction recognition) reliably contributed to participant classification and was strongly affected by coaching, especially Visual Reproduction recognition. More specifically, for every incorrectly recognized design, the probability of being misclassified as a true TBI survivor decreased by a considerable 63.5%; no other index had such a powerful influence over classification. It is worth noting that the large effect on probability is in part due to the small range of possible scores on this measure, but at the same time, is also due to the fact that unsuccessful simulators performed *disproportionately* poorly on this measure in comparison to bona fide TBI survivors.

The predictor with the second greatest influence was the Reliable Digit Span (RDS), with a 49% reduction in the probability of being classified as having a TBI for every 1-unit drop in the RDS. Interestingly, Verbal Paired Associates recognition produced an odds ratio that indicates that for every incorrectly identified item, the probability of being misclassified as having a TBI *increases* by 23%. This is a paradoxical finding, and it is inconsistent with the means for each group, which reflect worse performance among the unsuccessful simulators than bona fide TBI. In evaluating the base rate of the group mean for this variable, which fell in the 10^{th} to 16^{th} percentile, it becomes clear that TBI survivors performed poorly on this measure in comparison to the normative sample.

Although the present study offers support for the ACS method using these five indices as predictors of group membership simultaneously, additional models were explored due to the potential for excessive overlap among the predictors and the fact that not all variables are obtained from the WMS-IV. These models not only achieved comparable diagnostic efficiency, they utilized only indices from the WMS-IV and thus do not require administration of additional tests or use of the ACS package. The best fit of these models that contained only unique predictors without evidence of excessive overlap was a three-variable model dubbed the “Revised-ACS” model. Predictors in this model included the Brief Cognitive Status Exam (BCSE), a discriminability score from Verbal Paired Associates recognition, and Visual Reproduction recognition.

The Revised-ACS model demonstrated outstanding classification accuracy. It was the best fitting of the *a priori* specified models toward differentiation among good effort TBI and unsuccessful simulators, despite the fact that only two of the three individual predictors in this model were reliable in the sense that they offered stable predictions of participant classification. In the presence of the BCSE and Visual Reproduction recognition, the discriminability score did not contribute to classification accuracy. Even though refitting the model without the discriminability score may have improved the overall fit, to avoid artificial introduction of bias, it was retained as a predictor. As with the full version of the ACS, Visual Reproduction recognition had a dramatic influence on classification, though less so in this revised model than in the full ACS model, likely due to the additional presence of better fitting predictors. Similar to Visual Reproduction recognition, the BCSE influenced the probability of being classified in the same direction such that for every 1-unit drop in raw score on the BCSE, the probability of being classified as a TBI survivor is reduced; in this case by 13%. The weaker influence (e.g., smaller odds ratio) of the BCSE score in comparison to Visual Reproduction recognition is probably related to the wider range of possible scores for the BCSE that translates into greater variance.

Following development of the more complex models, individual predictors were evaluated including those used as part of the ACS and the Revised-ACS, refined analysis of Visual Reproduction recognition, and the longest span on the Spatial Span subtest. Of these indices, Visual Reproduction recognition was still the most powerful single predictor and demonstrated the highest classification accuracy. In comparison to the next best single-variable index – the BCSE – Visual Reproduction did a better job at identifying feigned cognitive impairment even though both were effective. Of note, *Occam's razor* (Thorburm, 1915) applies when comparing models using Bayesian methods: Also known as the *principle of parsimony*, it advises and rewards the simplest model among a set of otherwise equivalent solutions. Thus, although the Revised-ACS model was somewhat superior to Visual Reproduction recognition alone in terms of predictive accuracy, Bayesian fit statistics reward simplicity and penalize complexity, particularly with smaller to medium sample sizes (Hastie, Tibshirani, & Friedman, 2009); therefore, the Bayesian comparison equated Visual Reproduction recognition and the Revised ACS model. For theoretical or psychometric reasons, clinicians may choose either index, but because the ACS-Revised model does not require additional time or clinical resources in exchange for the increase in predictive accuracy, it is recommended as the more appropriate tool. However, in the event that the ACS measures are not administered, clinicians can feel confident in using the recognition score from Visual Reproduction.

Why individuals coached to feign memory impairment selected visual tasks as a target of simulation strategy is unclear. What is important to note, however, is that of the available recognition measures evaluated in the present study, Visual Reproduction recognition was the only one that does not utilize a forced-choice recognition paradigm; Verbal Paired Associates recognition and Logical Memory recognition both utilize a yes/no format. On Visual Reproduction recognition, respondents select the correct design from six foils, and perhaps the presentation of multiple response options allowed for individuals to be more creative in their responses and introduce more variability than a simply correct or incorrect response. In addition, there is greater probability for incorrect responses, though the same holds true for good effort participants. The other recognition memory task that has more than two response options was the Designs subtest, which demonstrated a medium effect based on standard interpretation guidelines; however, it was much smaller than the Visual Reproduction recognition effect size. In the present study, as well as in the Pearson ACS model, the Designs recognition subtest was not evaluated for its ability to detect cases of suboptimal effort. The reason Pearson excluded this test is unclear; however, it was excluded in the present study based on observations of a high proportion of random responding among participants in both groups.

One possibility is that the good effort TBI group performed particularly poorly on the Designs subtest, and its recognition component, which would attenuate the effect size and limit the utility of this measure as a measure of symptom validity. Even though only one of the two visual memory recognition measures was useful in differentiating between groups, it is hypothesized that visual memory, particularly drawings, are a common target of simulation strategy. This is not to suggest that verbal measures do not retain similar susceptibility; however, in the present study there was a more consistent effect of coaching on the visually based measures. Comparing effect sizes from the TOMM, which uses line drawings, to the WCT, a procedurally similar task using verbal stimuli, further supports this notion. This is most certainly a point worthy of further investigation in future research.

The other reliable predictor in the Revised ACS was the BCSE and when used in isolation of the other two predictors, it demonstrated good overall diagnostic efficiency. However, in an overall comparison, the complete Revised ACS model or Visual Reproduction recognition both are recommended over the BCSE alone. The performance of easily detected simulators on the BCSE was interesting, and at times comical. The allowance of open-ended responses afforded respondents even greater flexibility in their response style and this generated some rather outlandish answers. The most frequent errors were on the orientation portion of the measure, and the errors represent a potential source symptom validity screening. Although individuals with actual TBI are prone to errors in orientation, none of the good effort TBI survivors reported the current president of the United States as “Captain Kirk”. Confabulations of this nature are rare in bona fide neuropsychological disorders, although they do occur in conditions such as Wernicke's aphasia and Korsakoff's syndrome (Lezak et al., 2004). Another advantage of the BCSE is that it also offers several weighted scores for each individual cognitive domain assessed in the BCSE that could potentially offer greater predictive ability over the total raw score.

The BCSE is a good example of the public's general lack of awareness related to actual functioning after TBI. Many individuals asked to simulate TBI did very poorly on this measure, particularly among items related to orientation and visual naming. It appeared as though individuals asked to simulate assumed that TBI survivors are in a complete and permanent state of confusion and disorientation, which is typically not the case following an emergence from post-traumatic amnesia (Dikmen et al., 1995; Lezak et al., 2004).

One example within the present study in which score refinement proved to be useful was with Verbal Paired Associates recognition. The standard method of scoring simply provides the total number of correct responses, but with the release of the ACS came the ability to calculate separately the hard hits and easy hits, as well as the number of false positives and a discriminability score. Each of these indices were evaluated for the ability to differentiate among groups. The number of easy hits on its own was not sufficient to detect cases of simulated impairment. This was likely due to a very limited range (only four items are considered "easy"). However, very few of TBI survivors putting forth good effort missed any of the easy items, and nearly all got three of the four correct. As such, a score of 2 or less should raise suspicion. The number of false-positives was also not a reliable predictor, even though those feigning memory impairment committed more errors of this nature. The indices that did emerge as reliable were the number of hard hits and the discriminability score.

The number of hard hits was similar to the number of easy hits but with a larger range (there are 10 possible hard hit items), in that the TBI group performed relatively well and those in the simulator group did slightly worse. In the present study, this was the worst predictive index, and because of the low discriminability no cut off score was offered. Because other measures within this subtest offer better classification accuracy (e.g., Verbal Paired Associates recognition discriminability), clinicians are encouraged to use those instead of the number of hard hits, especially because such measures take the number of hard hits into account.

The Verbal Paired Associates recognition discriminability score was also calculated and proven to be a reliable predictor, despite rather poor differentiation between groups and being the least preferred index of those tested. Discriminability scores have been used in other recognition memory measures (e.g., CVLT-II) as well as prediction models for the detection of suboptimal effort (e.g., Wolfe et al., 2010). The reliability of this index appearing in measures of symptom validity is not surprising, considering the way in which such a score is calculated. It also highlights the strength of embedded measures and prediction models. To demonstrate good discrimination between targets requires a sophisticated response style and simulation strategy. Not only must examinees understand the way in which the value is determined, they must also possess exceptional working memory skills in order to keep track of their previous responses. Such mental efforts are likely to impede a person’s ability to concentrate on the task at hand *and* maintain their sophisticated approach to failing the test.

The visually based version of the Digit Span subtest also showed promise. Unfortunately, as was the case with the number of easy hits for Verbal Paired Associates recognition, the longest Spatial Span suffered from restriction of range, with the TBI group demonstrating a longest span that averaged only three designs. Despite such a restriction, the longest spatial span still was a far better fit than any of the Verbal Paired Associates recognition indices. Although the low level of performance of the TBI group suggests a high degree of difficulty for the task, the fact that simulators demonstrated even lower performance reinforces the susceptibility of visually based measures to feigned cognitive impairment. As with Visual Reproduction recognition, a single unit drop in longest span had dramatic influence over the probability of being classified as a simulator. Like many indices used in symptom validity testing (including the present study), the longest Spatial Span suffered from very low sensitivity; however, the specificity at a cutoff score of 2 was near perfect, with less than a 3% false-positive rate.

The only dedicated stand-alone SVT in the WMS-IV, the Word Choice Test, was surprisingly inefficient in comparison to some of the other measures (e.g., BCSE, Visual Reproduction recognition) at differentiating between cases of blatant feigned memory impairment and good effort TBI survivors. Using a cutoff score of 37 offered the best balance between false-positive and false-negative errors. This value essentially represents a level of chance performance and is consistent with research on similar measures (Horner, Bedwell, & Duong, 2006); however, in setting the criterion of passing performance at 37, more than 40% of test-takers withholding effort will go undetected. With such low sensitivity, if this index is adopted for clinical use, it is essential that it is used in conjunction with at least one other independent measure of symptom validity. Interestingly, the performance of the Word Choice Test improved considerably when differentiating between those who avoided detection and those readily identified as simulators, which will be discussed in detail in the following section.

A pattern that emerged in these findings is the relationship between the degree of flexibility in response options and the ability of a test to detect cases of blatant simulated memory impairment. In comparison to responses given by individuals putting forth good effort with a legitimate TBI, individuals simulating memory impairment appear to offer a greater degree of atypical responses. Although not formally analyzed in the present data, observations made during the course of data collection suggested that participants in the simulator sample made a very high number of intrusion errors, and more of these intrusions were often semantically and phonetically unrelated to any of the target words on the CVLT-II. This, too, is a promising avenue of future research.

One possibility is that this observation highlights a limitation of analog studies: Perhaps in an experimental setting, persons asked to fake cognitive impairment take a more cavalier approach than would be used by someone seeking remuneration; they have more fun with it because their performance has no bearing on the outcome. It may be that these individuals have little empirical knowledge about how legitimate memory impairment manifests itself, even though they know it is a common complaint, and thus they adopt unusual response styles. The counter to this argument, however, is that being in a car accident, suffering a concussion or some other form of mild injury, or involvement in active litigation does not simply increase awareness and understanding of memory loss (unless one retains a *really* unscrupulous lawyer).

<u>Section 6.2 – Hypothesis 2 (Unsuccessful Simulators vs. Successful Simulators)</u>

A prominent interest of the present project was to determine the differences in performance among those asked to simulate memory impairment that do so while avoiding detection (e.g., successful simulator) and those who are readily identified as a simulator (e.g., unsuccessful simulator). The approached used here, despite identical terminology, is different than what has previously been used in the literature (DenBoer & Hall, 2007). Although their study attempted to differentiate between successful and unsuccessful simulators, the criterion employed DenBoer and Hall (2007) was the influence of coaching on successful simulation. The findings of the present study suggest that this is an inadequate method of classification. In the current symptom validity literature, there is a lack of research investigating the differences among those who feign cognitive impairment. Most studies related to symptom validity testing use a well-validated external criterion in their classification procedures, such as failure of one or more specific SVT's, and all those meeting that criterion are thus considered to be withholding effort. The present findings clearly indicate that such practice is ill-advised because many individuals who reported feigning impairment in fact passed five out of six indices of effort.

One could argue that these individuals simply did not follow instructions and only endorsed that they did so on the debriefing questionnaire in order to appear as though the followed instructions. Considering that a handful of individuals chose not to follow instructions and honestly reported not doing so, as well as the fact that all participants in the study were clearly informed during the consent process that they would be compensated regardless of their performance, there is no apparent incentive for dishonesty in the debriefing.

Alternatively, it could be argued that these individuals did not follow instructions to feign memory impairment based on the fact that their scores were largely in the normal range and few successful simulators demonstrated impairment on the WMS-IV. This is not believed to be the case either. Within the present data, there are no other estimates of memory functioning when the individuals are putting forth maximal effort. Perhaps “Average” memory functioning for these participants is in fact the result of poor effort, and that under optimal effort circumstances, their memory functioning would be considered “High Average” or better. This is unlikely to be the most plausible explanation as it would then be expected that if this were the case – that Average memory function is the result of withholding effort for these individuals – estimates of intellectual ability would be considerably greater given the relationship between memory and IQ (Rapport, Axelrod, Theisen, Brines, Kalechstein, & Ricker, 1997), which they were not. Furthermore, in a clinical setting, an individual with Average memory functioning would not be considered as having suffered brain damaged.

A more plausible possibility is that these individuals adopted a strategy to appear brain injured so subtle that it was undetectable with the present measures (i.e., too subtle to appear impaired) or perhaps they attempted to appear brain injured in idiosyncratic ways that were not assessed by the effort indices (e.g., increased response latency), but still, the sum of their efforts on the present measures was not believable as appearing like TBI even though they were “successful” on the effort indices. Therefore, these individuals were classified as “successful simulators” based only on the logic that they were effective at avoiding detection on SVTs. Understanding how these individuals perform on cognitive testing is of great importance.

Unfortunately, research on this particular class of test-takers is extremely difficult. First and foremost, obtaining a sufficient sample size to comprehensively investigate patterns of performance beyond simple descriptive statements is difficult. In the present study, the base rate of successful simulation was only 28%. Although this seems like a high base rate, it is important to consider that they came from a group with a 100% incidence of exaggerated symptom presentation. In a clinical setting, assuming a base rate of 40%, out of every 100 people only 11 would go undetected as successful simulators. Even though this is considered a “rare” or “uncommon” phenomenon, this does not negate its clinical importance; most psychological disorders have lower base rates than this and are still areas of highly active research. In the present study, the sample size limited the number of predictors that could be effectively entered into a model. However, the individual predictors as well as a refined two-variable model were examined and their performance was compared to models used to discriminate good effort TBI from unsuccessful simulators. Although the full five-variable ACS model demonstrated good classification accuracy overall, this model cannot be considered valid until it is replicated with a larger sample. The same holds true for the Revised-ACS model examined in this context.

Another issue of difficulty, particularly in relation to Hypothesis 2, is that with such a small sample size and such a poor degree of performance of the unsuccessful simulator group in comparison to the successful simulator group, most prediction models using single variables are not much more informative than a simpler analysis of variance (ANOVA). With the exception being that logistic regression allows for creating ROC curves as well as diagnostic statistics, most of the single-variable prediction models were significant because of such discrepant performance. For example, of the single-variable predictors, the Word Choice test was best at differentiating between successful and unsuccessful simulators. Examining the average performance of each group, however, explains the finding, because the successful simulators averaged nearly perfect performance whereas the unsuccessful simulators averaged very poor performance. In other words, the finding primarily reflects the large discrepancy in performance; similar statements can be made regarding most of the individual predictors.

The successful simulators demonstrated a largely consistent pattern of reduced performance on subtests largely at the low end of the Average range. Specifically, there was a clear discrepancy on the verbal measures: the successful simulators performed considerably better on the verbal tasks than did any of the other groups. On the visual memory and working memory measures, however, their performances were far more consistent with the TBI group that put forth good effort; in other words, these individuals were effective at simulating bona fide TBI in that regard. This raises the question of how good our current methods of assessing symptom validity are if this group of individuals was able to avoid detection and still produce a pattern of performance that is essentially undifferentiated from actual TBI.

In this regard it is very noteworthy that although these “successful simulators” were able to avoid detection as fakers by putting forth adequate effort on the symptom validity indices, they did not achieve the ultimate goal of successful simulation of TBI; to do that required that they appeared brain injured on the cognitive evaluation. In fact, 28% of simulators successfully navigated the effort indices, but only a handful of those individuals (20% of all persons attempting to simulate) also successfully showed impairment on at least one of the WMS-IV memory indexes. These observations strengthen the notion that we can obtain valid estimates of persons purporting to have TBI-related cognitive deficits and that feigning is a complex and difficult task. Most individuals attempting to simulate TBI-related cognitive decline will be unsuccessful at two ends of the assessment spectrum: Too much and they are detected by indices of suboptimal effort, which identify the majority of simulators, too little and they appear too normal to have sustained a clinically meaningful TBI.

Interestingly, several models were a better fit for differentiating between cases of simulator success than for the detection of bona fide neurological injury. In particular, whereas Verbal Paired Associates recognition discriminability was one of the worst predictors for classification of TBI survivors, it was one of the best predictors at differentiating successful from unsuccessful simulators. Based on the pattern of performance of each group, it appears that the successful simulators were more sophisticated and careful in their responses than were the unsuccessful simulators, thus reinforcing the notion that appropriately balancing target recognition with false positives is difficult and its measurement is an effective index for detecting even well-executed symptom exaggeration. In the present data successful simulators generally showed better discriminability, essentially implying that they were more vigilant in balancing the number of hits and number of false-positives. Similarly, Verbal Paired Associates recognition discriminability was the only measure to differentiate between actual and well-simulated TBI, thus confirming the utility of this score. The number of hard-hits on Verbal Paired Associates recognition was also a better classification tool for differentiating successful versus unsuccessful simulators than for differentiating unsuccessful simulators from bona fide TBI.

Visual Reproduction recognition was again an excellent predictor, and when combined with the WCT, yielded the best model to differentiate successful versus unsuccessful simulators. The value of this model is that it contains performance on two separate measures from the WMS-IV, each of which showed good discriminability on their own, with one each from the visual domain and the verbal domain. In addition, one measure is a truly embedded measure, and the other is an independent stand-alone test. Overall, Visual Reproduction recognition showed a greater influence over classification in comparison to the WCT. Although this two-variable model was not initially examined as a tool to differentiate unsuccessful simulators from bona fide TBI, subsequent evaluation of the model in this context showed excellent ability at doing so. Of note, it was the best fitting model among those tested in section one even though it was slightly less efficient overall in classification accuracy in comparison to the Revised ACS model.

Section 6.3 – Good Effort TBI vs. Successful Simulators

Although this comparison was not specified as an original hypothesis, it seemed essential to address this question for both theoretical and clinical reasons. The ability to differentiate between successful simulators and bona fide neurological injury is of great importance. Although the models examined in this section suffer from the same issues and limitations as those examined in section 2, these exploratory analyses provided some important information.

Unfortunately, the only index that distinguished these two groups with any reliability was Verbal Paired Associates recognition discriminability, which is not surprising considering its widespread efficacy in previous models presented here as well as other research (Wolfe et al., 2010). In the present study, when used on its own it was not particularly effective, and in the absence of other measures of symptom validity its use is unadvisable. Its reliability in this context, however, strongly supports the notion that simulation and feigning memory impairment is a challenging task to do well.

One reason this index is so reliable in the detection of credible performance may be that it is a purely embedded measure, in that it is calculated based on patterns of performance within a measure as opposed to a simple performance indicator. With the exception of persons very well versed in psychometrics and/or test construction, few test takers are likely to be aware of this index, and as such, place little consideration to the factors contributing to its calculation (e.g., number of hits relative to the number of false-positives).

Even though other indices showed promise as potentially useful in combination with other measures, in the present study, no other indices reliably differentiated between successful simulators and actual TBI survivors putting forth good effort. Although indices may in fact exist to this end, short of a haphazard approach to variable selection, no such model was identified, and the present sample size precludes use of statistically driven model specification (e.g., Bayesian Model Averaging). Based on these observations, ongoing research in this area is certainly needed and of substantial importance.

Section 6.4 – Study Limitations

A major limitation of the present study is that it has been argued that analog studies of malingering are limited in the generalizability of their findings (e.g., Larrabee, 2007; Rogers, 1997b; Suhr & Gunstad, 2007), which has implications within the forensic arena. However, the increasing presence of neuropsychology within a litigating environment brings with it the increased potential for coaching, particularly by plaintiff attorneys seeking to help their clients secure external rewards. Given this possibility, there is a prominent need for empirical evidence documenting the patterns of performances observed in response to coaching of neurologically intact individuals (Coleman et al., 1998). Conducting a study that has a known group of individuals who are likely to be actual malingerers would represent the strongest design. The current study, however, certainly lessens the gap between a pure analog study and a known-groups design due to its inclusion of persons with verified TBI. Furthermore, in the development phases of an embedded SVT, analog studies can be particularly helpful as they involve minimal risk to participants and are well suited to follow-up with more robust designs such as a known-groups design.

Another limitation of the present study is the small sample size of the successful simulator group. With fewer than 20 participants in this group, the ability to evaluate reliable and stable prediction models was substantially limited to two indices per model. Although standards do exist that have suggested that models can be fit with a subject to predictor ratio of 6 to 1, this is at the more liberal end of the cutoff. Much greater confidence can be placed in a model with at least a 10 to 1 ratio. Despite this shortcoming, the present study certainly provides for a foundation upon which further research can be built. This study has successfully identified the predictive abilities of single variables that can potentially be combined far more effectively with a larger sample.

Section 6.5 – Study Conclusions and Future Directions

In conclusion, several themes can be taken from the present study. First and foremost, the findings from the present study strongly support the use of multiple measures of symptom validity interspersed throughout testing. Under no circumstances should a clinician base a conclusion – especially one with such high risk associated with a false positive – on a single piece of data, and the present results clearly support this principle. In addition, this study has also provided firm support for the use of the WMS-IV in detecting cases of feigned cognitive impairment and offered several methods to effectively do so. The present study has also shown just how difficult it is to simulate a TBI in a way that is credible.

The present has also generated a number of avenues to pursue in future research endeavors. One such pursuit for effort testing should evaluate more refined pattern analysis of individual scores and performances across multiple measures. The present study has demonstrated that there are differences in approaches among individuals feigning cognitive impairment, and some are more effective at avoiding detection than others. For example, some individuals may adopt very unique strategies to simulate impairment and the present battery of SVTs available to the clinician may not be sufficient to detect these strategies. As such, future research should attempt to both understand these unique approaches and to develop measures to detect them.

Also of interest along the lines of pattern performance is to further explore the notion of targeted simulation. In the present study there appeared to be a trend in simulators to target visually based measures as the target of their strategies, and gaining more insight into the reliability of this phenomenon in other samples may provide direction towards development of future methods of detecting feigned cognitive impairment.

An additional point of future research would be to engage in error analysis of individuals feigning impairment. For example, previous research has evaluated the semantic relationship of intrusion errors on VPA to the targets and stimuli (Rapport, Axelrod, & Mansharamani, 1995). Observations made during the course of data collection in the present study also suggested that individuals with TBI generated more intrusions that were of high semantic relatedness in comparison to the simulator sample. Alternatively, a similar approach would be to evaluate base rate data of error responses. A high number of intrusions with low or non-existent semantic associations may be indicative of an intentional response bias.

One of the most intriguing portions of the present study was the sample of TBI survivors who put forth less than optimal effort. In a clinical setting, the results of these individuals would be deemed invalid and the individuals would have possibly been labeled as “malingering”. However, there is no readily apparent motivation for them to respond in such a way, raising the possibility that these individuals were not necessarily exaggerating their impairments, but that they simply were not putting forth maximal effort. If this is the case, it highlights the ability of “effort” tests to detect less than full effort. However, it also reaffirms the notion that a failure of on an SVT does not allow for differentiation between less than full effort and a negative response bias. Thus, it is recommended that future research investigate the patterns of cognitive performance that differentiate individuals who are simply not putting forth sufficient effort from those who are actively malingering in a clinical setting.

APPENDIX A

TABLES

Table 1. Slick et al. (1999) Diagnostic Criteria for Malingering

	Criteria	Definite	Probable	Possible
Criteria A.	*Presence of a substantial external incentive*	**X**	**X**	**X**
Criteria B.	*Evidence from neuropsychological testing*			
1.	*Definite negative response bias (Below chance performance (p < .05) on one or more forced-choice measures)*	**X**		
2.	*Probable response bias (Performance on one or more well-validated psychometric tests or indices designed to measure exaggeration or fabrication of cognitive deficit is consistent with feigning)*		†	
3.	*Discrepancy between test data and known patterns of brain functioning*		†	
4.	*Discrepancy between test data and observed behavior*		†	
5.	*Discrepancy between test data and reliable collateral reports*		†	
6.	*Discrepancy between test data and documented background history*		†	
Criteria C.	*Evidence from self report*			
1.	*Self-reported history history is discrepant with documented history*		†	‡
2.	*Self-reported symptoms are discrepant with known patterns of brain functioning*		†	‡
3.	*Self-reported symptoms are discrepant with behavioral observations*		†	‡
4.	*Self-reported symptoms are discrepant with information obtained from collateral informants*		†	‡
5.	*Evidence of exaggerated or fabricated psychological dysfunction*		†	‡
Criteria D.	*Behaviors meeting necessary criteria from groups B or C are not fully accounted for by psychiatric, neurological, or developmental factors*	**X**	**X**	‡

† Two or more criteria from B2 - B6 or one criteria from B2 - B6 and one or more from C1 - C5

‡ One or more from C1 - C5 and criteria D *or* Criteria for Definite or Probable is met with the exception of criteria D

Note: Copied from Slick, Sherman, & Iverson (1999)

Table 2. *Demographic Data.*

	Good Effort TBI (*n* = 41)		Suspect Effort TBI (*n* = 19)		Successful Simulator (*n* = 18)[1]		Unsuccessful Simulator (*n* = 41)[2]	
Variable	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
Months since injury	108.7	(75.0)	119.5	(73.2)	*NA*	-	*NA*	-
Glasgow Coma Scale	9.3	(3.9)	9.7	(4.2)	*NA*	-	*NA*	-
Age (years)	43.0	(12.3)	46.0	(11.6)	49.7	(8.2)	42.0	(12.1)
Education (years)	12.0	(2.3)	12.0	(2.2)	13.3	(2.4)	12.9	(2.0)
WTAR FSIQ	86.2	(9.9)	82.1	(6.6)	96.2	(13.4)	92.8	(11.9)
Gender (% men)	90%		100%		83%		80%	
Race/Ethnicity								
African American	76%		95%		61%		73%	
Caucasian	22%		5%		39%		27%	
Hispanic	2%		0%		0%		0%	

Note. WTAR = Wechsler Test of Adult Reading; FSIQ = Full Scale IQ.

1. Coached to feign cognitive impairment and failed fewer than two effort measures; WTAR administered prior to coaching instructions.
2. Coached to feign cognitive impairment and failed two or more effort measures; WTAR administered prior to coaching instructions.

Table 3. *Wechsler Memory Scale-IV Index Scores by Group.*

Variable	Good Effort TBI (*n* = 41)			Suspect Effort TBI (*n* = 19)			Successful Simulator (*n* = 18)[1]			Unsuccessful Simulator (*n* = 41)[2]			
	M	*(SD)*	*% Impaired*	*M*	*(SD)*	*% Impaired*	*M*	*(SD)*	*% Impaired*	*M*	*(SD)*	*% Impaired*[3]	*Cohen's d*
Auditory Memory	79.8	(16.9)	73.2%	64.3	(13.9)	94.1%	93.1	(13.2)	27.8%	73.2	(13.6)	82.9%	0.44
Visual Memory	85.5	(11.6)	48.8%	73.8	(15.2)	75.0%	86.7	(14.0)	55.6%	73.3	(15.7)	80.5%	0.89
Visual Working Memory	83.0	(11.1)	78.0%	70.9	(9.3)	88.9%	87.3	(12.0)	44.4%	66.8	(15.8)	90.2%	1.20
Immediate Memory	78.7	(14.3)	68.3%	63.1	(14.0)	94.1%	87.3	(14.8)	50.0%	68.8	(16.2)	82.9%	0.66
Delayed Memory	82.5	(13.5)	61.0%	68.7	(12.8)	93.8%	89.7	(12.9)	44.4%	72.0	(13.6)	87.8%	0.78

Note. Effect sizes calculated between Unsuccessful Simulator and Good Effort TBI samples using a pooled standard deviation.

1. Coached to feign cognitive impairment and failed fewer than two effort measures; reported faking as instructed on debriefing survey.
2. Coached to feign cognitive impairment and failed two or more effort measures; reported faking as instructed on debriefing survey.
3. % Impaired = percent of scores 1 SD or more below the mean.

Table 4. *Wechsler Memory Scale-IV Substituted Scores by Group.*

	Good Effort TBI (*n* = 41)		Suspect Effort TBI (*n* = 19)		Successful Simulator (*n* = 18)[1]		Unsuccessful Simulator (*n* = 41)[2]		Cohen's d
Variable[3]	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	
Auditory Memory	80.3	(17.4)	68.6	(14.2)	93.0	(10.9)	71.3	(15.7)	0.55
Immediate Memory	79.3	(14.8)	66.5	(12.6)	87.7	(14.0)	69.4	(17.6)	0.62
Delayed Memory	81.4	(14.5)	70.3	(13.5)	89.1	(12.4)	70.1	(13.5)	0.82
Verbal Paired Associates 1	7.4	(2.7)	5.5	(2.4)	9.1	(2.8)	6.2	(3.1)	0.42
Verbal Paired Associates 2	7.3	(3.7)	5.2	(3.3)	8.9	(2.9)	5.8	(3.2)	0.44

Note. Effect sizes calculated between Unsuccessful Simulator and Good Effort TBI samples using a pooled standard deviation.

1. Coached to feign cognitive impairment and failed fewer than two effort measures; reported faking as instructed on debriefing survey.
2. Coached to feign cognitive impairment and failed two or more effort measures; reported faking as instructed on debriefing survey.
3. Based on substitution of CVLT-II Trial 1-5 T score and Long-delay free-recall Z scores for Verbal Paired Associates I and II, respectively.

Table 5. *Additional Wechsler Memory Scale-IV Scores by Group.*

	Good Effort TBI (*n* = 41)		Suspect Effort TBI (*n* = 19)		Successful Simulator (*n* = 18)[1]		Unsuccessful Simulator (*n* = 41)[2]		Cohen's d
Variable	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	
Designs 1									
Content	7.8	(3.0)	6.6	(3.1)	8.4	(2.2)	6.9	(2.7)	0.32
Spatial	8.0	(2.4)	6.4	(2.9)	7.9	(2.7)	7.0	(2.4)	0.42
Designs 2									
Content	9.0	(2.1)	7.1	(2.3)	8.8	(2.5)	7.4	(2.8)	0.65
Spatial	9.1	(2.4)	8.8	(2.6)	8.7	(1.9)	8.0	(2.1)	0.49
Logical Memory Recognition	21.8	(4.2)	19.8	(4.0)	24.2	(4.0)	20.4	(4.6)	0.32
Verbal Paired Associates Recognition	35.3	(4.5)	29.6	(6.0)	38.1	(2.7)	31.0	(6.9)	0.75
Visual Reproduction Recognition	5.6	(1.1)	3.6	(1.3)	5.6	(1.5)	2.8	(1.9)	1.83
Designs Recognition	13.7	(2.1)	12.1	(3.0)	12.7	3.1	11.8	(3.9)	0.57
Word Choice	46.2	(4.6)	41.6	(7.1)	48.1	(3.4)	35.1	(11.7)	1.26
Brief Cognitive Status Exam	48.3	(7.8)	40.1	(10.4)	45.6	(12.2)	31.1	(13.9)	1.55

1. Coached to feign cognitive impairment and failed fewer than two effort measures; reported faking as instructed on debriefing survey.
2. Coached to feign cognitive impairment and failed two or more effort measures; reported faking as instructed on debriefing survey.

Table 6. *Effort Indicators Performance by Group.*

	Good Effort TBI (n = 41)			Suspect Effort TBI (n = 19)			Successful Simulator (n = 18)[1]			Unsuccessful Simulator (n = 41)[2]			
Variable	*M*	*(SD)*	*Failure Rate*[3]	*M*	*(SD)*	*Failure Rate*	*M*	*(SD)*	*Failure Rate*	*M*	*(SD)*	*Failure Rate*	*Cohen's d*
Test of Memory Malingering													
Trial 1	45.4	(4.4)	*NA*	41.2	(5.8)	*NA*	46.8	(3.7)	*NA*	32.0	(10.0)	*NA*	1.76
Trial 2	49.3	(1.9)	2%	46.6	(4.7)	26%	49.7	(0.8)	0%	34.1	(12.8)	76%	1.68
Reliable Digit Span	8.6	(1.9)	12%	6.3	(1.5)	47%	8.4	(1.4)	6%	5.6	(2.8)	51%	1.27
CVLT-II Forced Choice	15.8	(0.6)	3%	15.2	(1.6)	17%	15.9	(0.3)	0%	13.0	(3.2)	59%	1.21
CVLT-II 3-Variable Model	0.45	(0.2)	42%	0.56	(0.2)	72%	0.30	(0.2)	11%	0.60	(0.2)	71%	0.76

Note. Effect sizes calculated between Unsuccessful Simulator and Good Effort TBI samples using a pooled standard deviation.

1. Coached to feign cognitive impairment and failed fewer than two effort measures; reported faking as instructed on debriefing survey.
2. Coached to feign cognitive impairment and failed two or more effort measures; reported faking as instructed on debriefing survey.
3. Failure Criteria: TOMM Trial 2 < 45; Reliable Digit Span ≤ 6; CVLT-II Forced Choice ≤ 14; CVLT-II 3-Variable Predicted Probability > 0.5.

Table 7a. *Green's Effort Measures Performance by Group: Non-Verbal Medical Symptom Validity Test.*

	Good Effort TBI (*n* = 25)		Suspect Effort TBI (*n* = 14)		Successful Simulator (*n* = 16)[1]		Unsuccessful Simulator (*n* = 39)[2]		
Variable	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*Cohen's d*
Immediate Recall	99.6	(2.0)	97.5	(7.0)	100.0	(0.0)	77.4	(28.5)	1.01
Delayed Recall	95.6	(8.2)	83.2	(17.4)	94.4	(8.9)	65.8	(27.9)	1.35
Consistency	95.2	(8.5)	82.9	(16.6)	94.4	(8.9)	73.5	(19.4)	1.37
Delayed Recognition Archetypes	83.2	(14.9)	68.9	(18.9)	86.9	(12.0)	63.3	(23.0)	1.00
Delayed Recognition Variations	96.4	(8.1)	85.7	(16.5)	94.4	(8.1)	57.2	(31.7)	1.57
Paired Associates	99.2	(4.0)	92.1	(19.3)	100.0	(0.0)	81.3	(23.2)	1.23
Free Recall	60.6	(21.6)	47.9	(21.1)	61.3	(20.6)	42.6	(18.9)	0.89
Failure Rate[3]	16%		86%		19%		85%		

Note. Effect sizes calculated between Unsuccessful Simulator and Good Effort TBI samples using a pooled standard deviation.
1. Coached to feign cognitive impairment and failed fewer than two effort measures; reported faking as instructed on debriefing survey.
2. Coached to feign cognitive impairment and failed two or more effort measures; reported faking as instructed on debriefing survey.
3. Failure Criteria: Either the mean of the IR, DR, CNS, DRA, DRV and PA is 90% or below OR the mean of DR, CNS, DRA and DRV is below 88%.

Table 7b. *Green's Effort Measures Performance by Group: Medical Symptom Validity Test.*

	Good Effort TBI (*n* = 38)		Suspect Effort TBI (*n* = 19)		Successful Simulator (*n* = 17)[1]		Unsuccessful Simulator (*n* = 41)[2]		
Variable	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*Cohen's d*
Immediate Recall	98.2	(4.1)	93.2	(7.7)	99.4	(1.7)	69.1	(25.3)	1.60
Delayed Recall	98.2	(4.3)	86.3	(17.6)	99.1	(2.0)	66.8	(25.6)	1.70
Consistency	97.1	(6.2)	84.7	(18.4)	98.5	(2.9)	71.3	(20.2)	1.72
Paired Associates	95.8	(9.3)	69.5	(26.1)	95.9	(7.9)	59.5	(29.2)	1.67
Free Recall	59.2	(19.3)	35.8	(19.9)	70.3	(12.9)	40.9	(19.6)	0.95
Failure Rate[3]	5%		47%		0%		78%		

Note. Effect sizes calculated between Unsuccessful Simulator and Good Effort TBI samples using a pooled standard deviation.
1. Coached to feign cognitive impairment and failed fewer than two effort measures; reported faking as instructed on debriefing survey.
2. Coached to feign cognitive impairment and failed two or more effort measures; reported faking as instructed on debriefing survey.
3. Failure Criteria = At least one of the IR, DR, or CNS scores is at or below 85%.

Table 8. *Test performance on ACS measures when grouped without use of the Reliable Digit Span as a criterion.*

	Good Effort TBI (*n* = 45)		Unsuccessful Simulators (*n* = 39)		
Variable	*M*	*(SD)*	*M*	*(SD)*	*Cohen's d*
Reliable Digit Span	8.2	(2.2)	5.6	(2.9)	1.03
Logical Memory Recognition	21.5	(4.2)	20.6	(4.4)	0.21
Verbal Paired Associates Recognition	34.8	(5.0)	30.8	(7.0)	0.67
Visual Reproduction Recognition	5.6	(1.2)	2.8	(1.9)	1.81
Word Choice	46.5	(4.4)	34.7	(11.8)	1.38

Table 9. *Hypothesis 1 logistic regression model summary statistics (Good Effort TBI vs. Unsuccessful Simulators).*

	B	*p*	Odds Ratio	95% C.I.
Full ACS Model:				
Word Choice Test	-0.218	.007	0.804	0.687 – 0.941
Reliable Digit Span	-0.670	.001	0.512	0.345 – 0.760
Logical Memory Recognition	-0.075	.572	0.928	0.715 – 1.204
Verbal Paired Associates Recognition	0.269	.025	1.309	1.035 – 1.656
Visual Reproduction Recognition	-1.009	.001	0.365	0.203 – 0.655
Education	0.211	.268	1.234	0.850 – 1.792
Constant	8.426	.034		
ACS Revised Model:				
Brief Cognitive Status Exam	-0.140	.002	0.870	0.795 – 0.951
VPA Recognition Discriminability	1.946	.502	7.003	0.024 – 2053.194
Visual Reproduction Recognition	-0.833	.000	0.435	0.275 – 0.688
Constant	1.931	.001		

Table 10. *Hypothesis 2 logistic regression model summary statistics (Successful Simulators vs. Unsuccessful Simulators).*

	B	*p*	Odds Ratio	95% C.I.
ACS 4-Variable:				
Reliable Digit Span	-0.567	0.05	0.567	0.32 – 1.00
Logical Memory Recognition	0.097	0.54	1.102	0.81 – 1.51
Verbal Paired Associates Recognition	-0.231	0.23	0.793	0.55 – 1.15
Visual Reproduction Recognition	-0.639	0.03	0.528	0.30 – 0.92
Constant	13.73	0.02		–
ACS 2-Variable:				
Word Choice Test	-0.269	0.02	0.76	0.62 – 0.95
Visual Reproduction Recognition	-0.565	0.02	0.57	0.36 – 0.90
Constant	15.306	0.00		

Table 11. *Prediction model summary table.*

	Traumatic Brain Injury vs. Unsuccessful Simulator (Hypothesis 1)		Unsuccessful Simulator vs. Successful Simulator (Hypothesis 2)	
	AUC	BIC	AUC	BIC
Word Choice Test	0.830	-271.75	0.897	-188.81
Brief Cognitive Status Exam	0.874	-279.10	0.813	-174.81
Visual Reproduction Recognition	0.878	-284.44	0.857	-183.96
Verbal Paired Associates Recognition - Hard Hits	0.701	-250.32	0.793	-172.82
Verbal Paired Associates Recognition - Discriminability	0.718	-248.59	0.864	-180.23
Longest Spatial Span	0.750	-260.45	0.808	-178.27
Revised ACS Model	0.929	-284.84	0.911	-179.79
2-Variable ACS Model	0.907	-286.36	0.935	-191.62
Full Pearson ACS Model	0.950	-	0.913	-

Note. BIC values for the Full Pearson ACS model cannot be compared to the other models presented, because the models were fitted using different samples; AUC statistics may be compared.

APPENDIX B

FIGURES

Figure 1. WMS-IV subtest performance profile by group

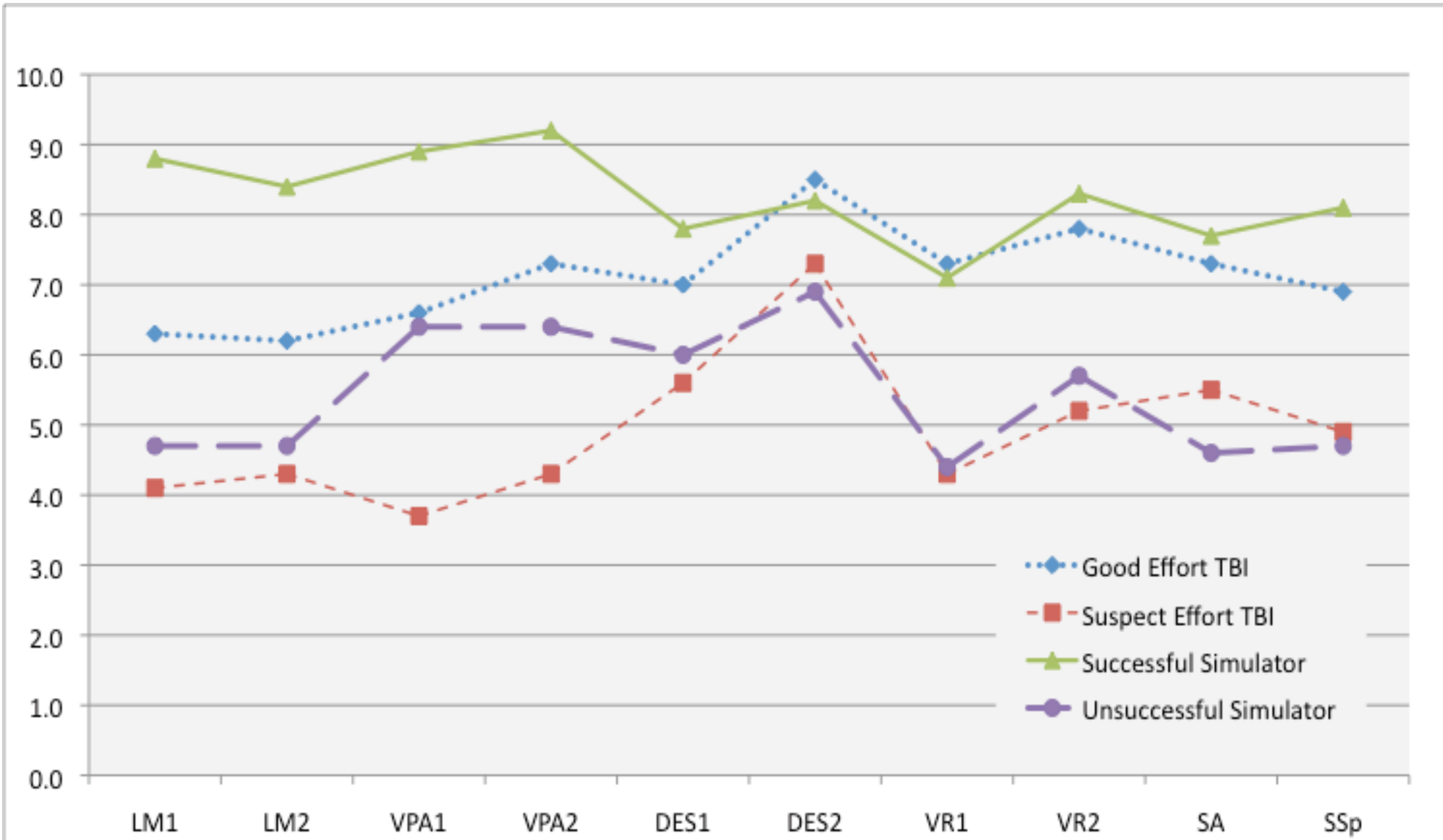


Figure 2. ROC Curve for Pearson ACS Full Model (Good Effort TBI vs. Unsuccessful Simulators)

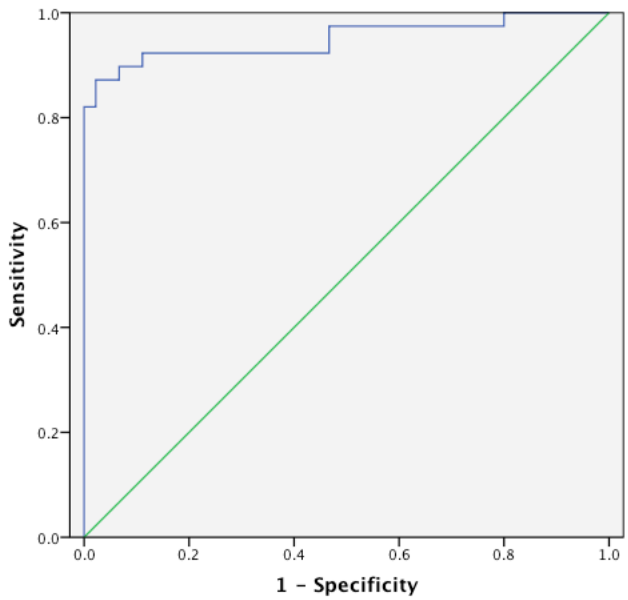


Figure 4. ROC Curve for Brief Cognitive Status Exam (Good Effort TBI vs. Unsuccessful Simulators)

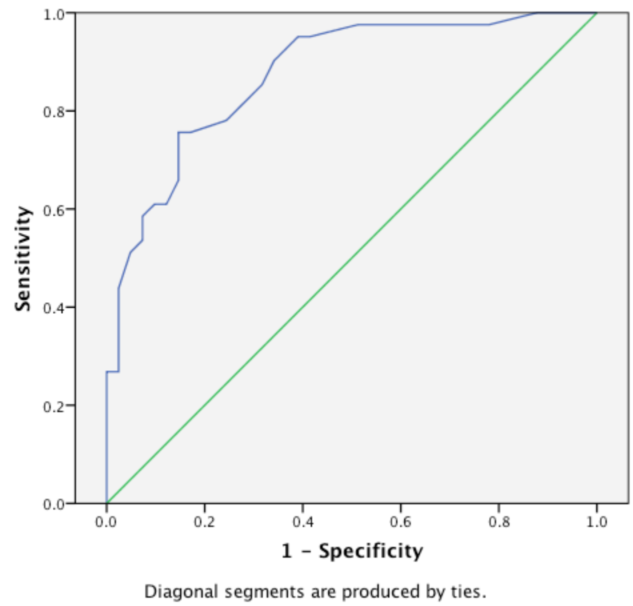


Figure 3. ROC Curve for WCT Model (Good Effort TBI vs. Unsuccessful Simulators)

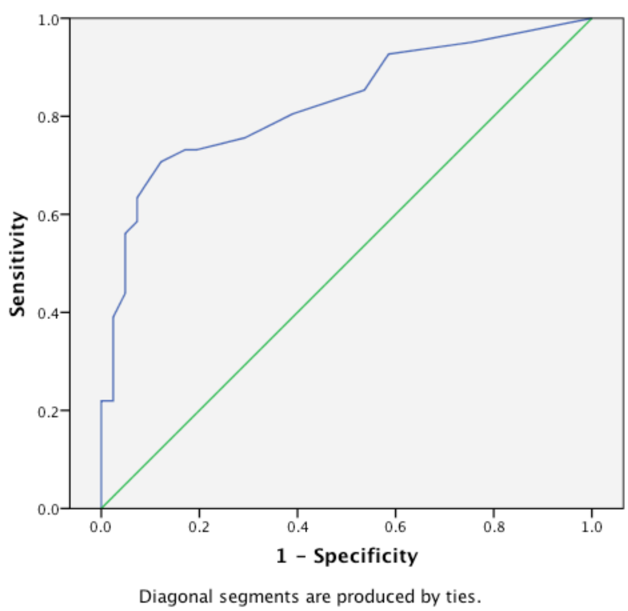


Figure 5. ROC Curve for Visual Reproduction Recognition (Good Effort TBI vs. Unsuccessful Simulators)

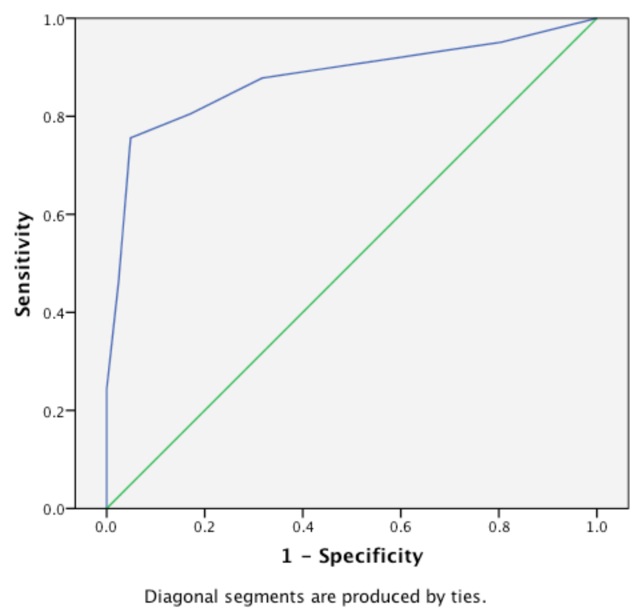


Figure 6. ROC Curves for Verbal Paired Associates recognition variables (Good effort TBI vs. Unsuccessful Simulator)

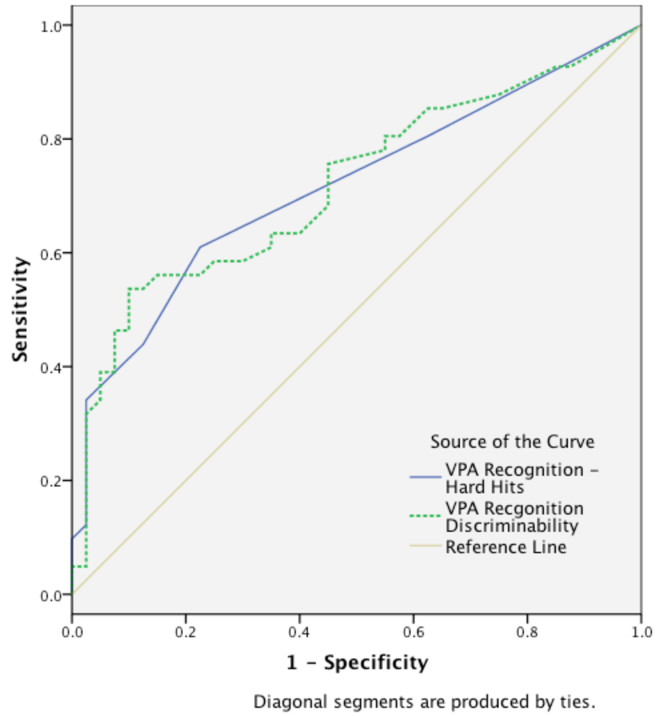


Figure 7. ROC Curve for the ACS Revised model (Good effort TBI vs. Unsuccessful Simulator)

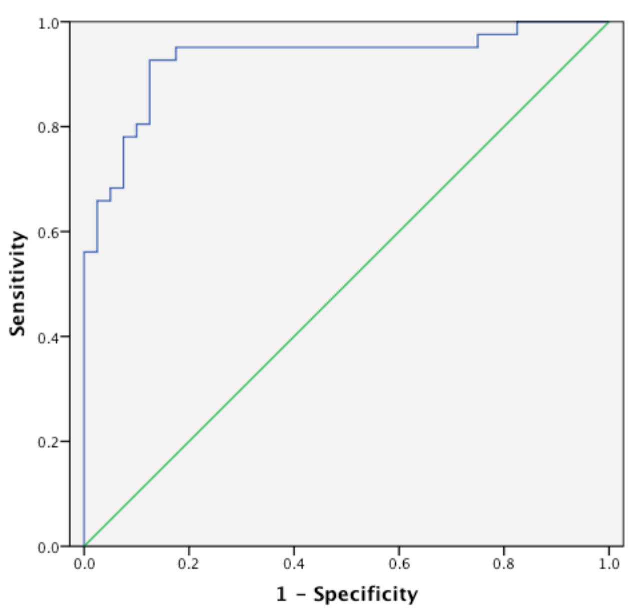


Figure 8. ROC curve for the ACS model without the WCT (Successful Simulator vs. Unsuccessful Simulator)

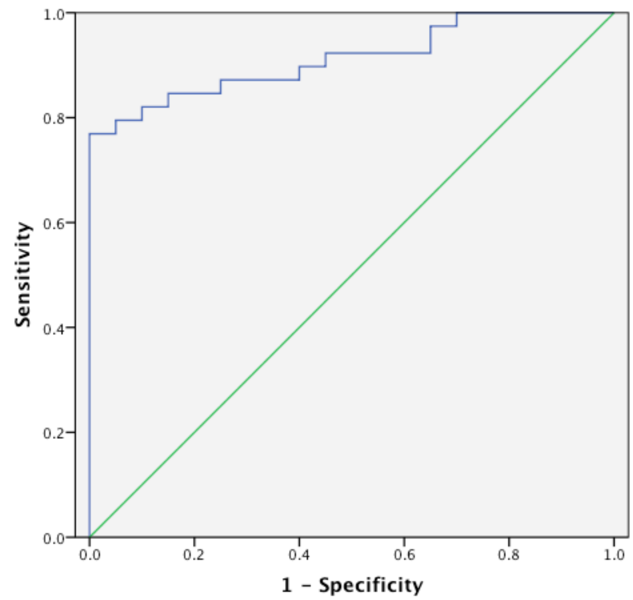


Figure 9. ROC Curve for the Word Choice test only (Successful Simulator vs. Unsuccessful Simulator)

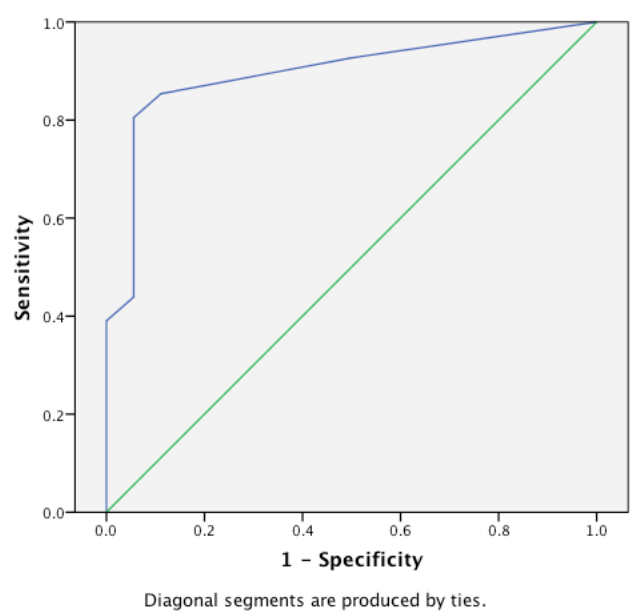


Figure 10. ROC Curve for the Brief Cognitive Status Exam (Successful Simulator vs. Unsuccessful Simulator)

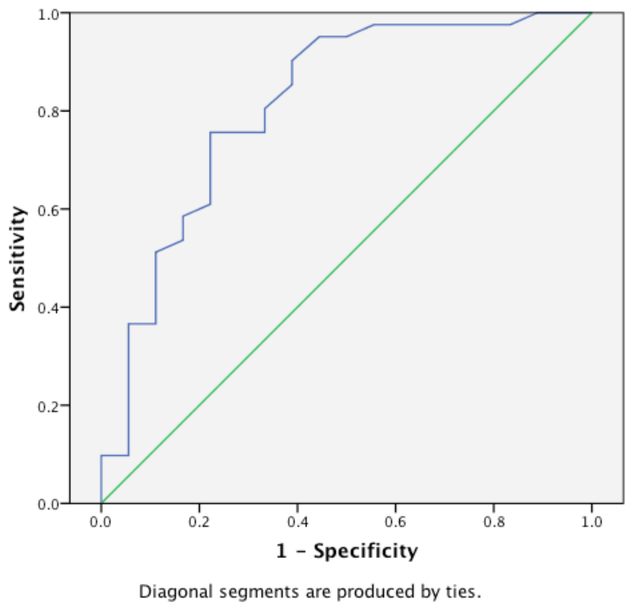


Figure 12. ROC Curve for 2-Variable ACS Model (Successful Simulator vs. Unsuccessful Simulator)

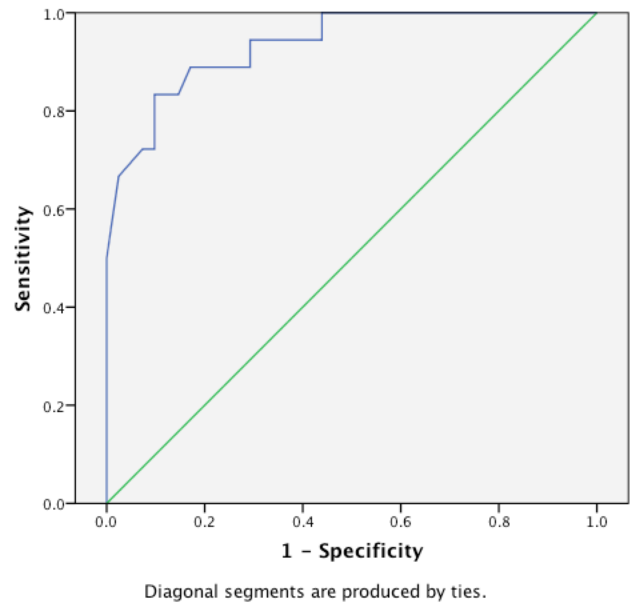
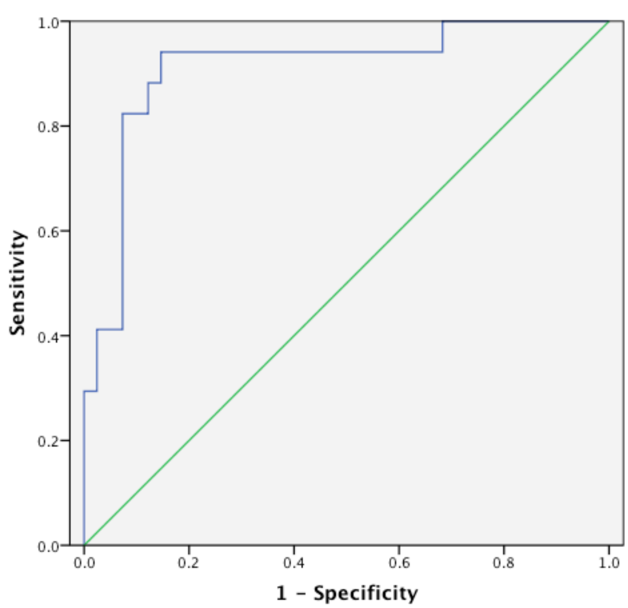


Figure 11. ROC Curve for ACS-Revised (Successful Simulator vs. Unsuccessful Simulator)



APPENDIX C

SIMULATOR ACCIDENT SCENARIO (ADAPTED FROM: TOMBAUGH, 1997)

"In this study you will be asked to complete several tasks that are often used to measure a variety of changes that occur in people who have brain damage. As you take each test, I would like you to assume the role of someone who has experienced some brain damage from a car accident.

Pretend that you were involved in a head-on collision. You hit your head against the windshield and were unconscious for 15 minutes. You were hospitalized overnight for observation and then released. Gradually, over the past few months, you have started to feel normal again. However, your lawyer has informed you that you may obtain a larger settlement from the court if you look like you are still suffering from brain damage. Therefore, you should pretend that the symptoms have persisted and that they still significantly interfere with your life.

As you portray the above person, try to approach each test as you imagine this person would respond if the individual had been given the same instructions from his or her lawyer. Perform on the tests in such a way as to convince the examiner that you are truly brain damaged, keeping in mind that settlement monies depend upon your being diagnosed as cognitively impaired. Also be aware that having a lawsuit pending often raises the suspicion that people may try to exaggerate their difficulties. This means that your impairments resulting from the head injury must be believable. Major exaggerations, such as not being able to do anything, remembering absolutely nothing, or failing to respond are easy to detect."

APPENDIX D

ELIGIBILITY CRITERIA FOR INCLUSION IN THE SOUTHEASTERN MICHIGAN TRAUMATIC BRAIN INJURY SYSTEM

In order to be eligible SEMTBIS participants, persons must have suffered a moderate to severe TBI (post-traumatic amnesia > 24 hours, loss of consciousness > 30 minutes, Glasgow Coma Scale score < 13 upon presentation to the emergency department, or intracranial neuroimaging abnormalities), must have been at least 16 years of age at the time of injury, received acute care at a designated model system site within 72 hours after injury, been directly transferred to a model system inpatient rehabilitation unit, and given informed consent.

REFERENCES

Ahmed, S., Bierley, R., Sheikh, J. I., & Date, E. S. (2000). Post-traumatic amnesia after closed head injury: a review of the literature and some suggestions for further research. *Brain Injury, 14*(9), 765-780.

AMA, E.-B. M. W. G. (1992). Evidence-based medicine: a new approach to teaching and the practice of medicine. *Journal of the American Medical Association, 268*(17), 2420 - 2425.

APA. (2002). Ethical Principles of Psychologists and Code of Conduct Retrieved January 6, 2008, from http://www.apa.org/ethics/code2002.html

APA. (2006). Evidence-based practice in psychology. *American Psychologist, 61*(4), 271 - 285.

Association, A. P. (2000). *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition, Text Revision ed.). Washington D.C.: American Psychitric Association.

Axelrod, B. N. (2008). *Cognitive and Emotional Differences among Veterans with Good and Poor Effort*. Paper presented at the Annual Meeting of the National Academy of Neuropsychology, New York, NY.

Axelrod, B. N., Barlow, A., & Paradee, C. (2010). Evaluation of the WMS-III Rarely Missed Index in a Naive Clinical Sample. *Clinical Neuropsychologist, 24*(1), 95-102.

Axelrod, B. N., Fichtenberg, N. L., Millis, S. R., & Wertheimer, J. C. (2006). Detecting incomplete effort with digit span from the Wechsler Adult Intelligence Scale - Third edition. *Clinical Neuropsychologist, 20*(3), 513-523.

Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various digit span scores in the detection of suspect effort. *Clinical Neuropsychologist, 20*(1), 145-159.

Baddeley, A. D. (1997). *Human Memory: Theory and Practice* (Revised ed.). East Sussex, UK: Psychology Press.

Baddeley, A. D. (2002). The psychology of memory. In A. D. Baddeley, M. D. Kopelman & B. A. Wilson (Eds.), *The handbook of memory disorders* (2nd ed.). West Sussex, UK: Wiley.

Bauer, L., & McCaffrey, R. J. (2006). Coverage of the Test of Memory Malingering, Victoria Symptom Validity Test, and word memory test on the Internet: Is test security threatened? *Archives of Clinical Neuropsychology, 21*(1), 121-126.

Binder, L. M. (1986). Persisting symptoms after mild head-injury - a review of the postconcussive syndrome. *Journal of Clinical and Experimental Neuropsychology, 8*(4), 323-346.

Binder, L. M. (1997). A review of mild head trauma .2. Clinical Implications. *Journal of Clinical and Experimental Neuropsychology, 19*(3), 432-457.

Binder, L. M., & Rohling, M. L. (1996). Money matters: A meta-analytic review of the effects of financial incentives on recovery after closed-head injury. *American Journal of Psychiatry, 153*(1), 7-10.

Binder, L. M., Rohling, M. L., & Larrabee, G. J. (1997). A review of mild head trauma .1. Meta-analytic review of neuropsychological studies. *Journal of Clinical and Experimental Neuropsychology, 19*(3), 421-431.

Boake, C., Freeland, J. C., Ringholz, G. M., Nance, M. L., & Edwards, K. E. (1995). Awareness of memory loss after severe closed-head injury. *Brain Injury, 9*(3), 273-283.

Boake, C., Millis, S. R., High, W. M., Delmonico, R. L., Kreutzer, J. S., Rosenthal, M., Sherer, M., & Ivanhoe, C. B. (2001). Using early neuropsychologic testing to predict long-term productivity outcome from traumatic brain injury. *Archives of Physical Medicine and Rehabilitation, 82*(6), 761-768.

Bortnik, K. E., Boone, K. B., Marion, S. D., Amano, S., Ziegler, E., Victor, T. L., & Zeller, M. A. (2010). Examination of various WMS-III logical memory scores in the assessment of response bias. *Clinical Neuropsychologist, 24*(2), 344-357.

Brooks, D. N. (1976). Wechsler Memory Scale performance and its relationship to brain damage after severe closed head injury. *J Neurol Neurosurg Psychiatry, 39*(6), 593-601.

Brown, A. W., Malec, J. F., McClelland, R. L., Diehl, N. N., Englander, J., & Cifu, D. X. (2005). Clinical elements that predict outcome after traumatic brain injury: A prospective multicenter recursive partitioning (decision-tree) analysis. *Journal of Neurotrauma, 22*(10), 1040-1051.

Bush, S. S., Ruff, R. M., Troster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity - NAN policy & planning committee. *Archives of Clinical Neuropsychology, 20*(4), 419-426.

Carone, D. A., & Turk, M. A. (2008). Validation of the medical symptom validity test (MSVT) in children with moderate to severe brain damage/dysfunction. [Meeting Abstract]. *Clinical Neuropsychologist, 22*(3), 440-440.

Cicerone, K. D., & Kalmar, K. (1995). Persistent postconcussion syndrome - the structure of subjective complaints after mild traumatic brain injury. *Journal of Head Trauma Rehabilitation, 10*(3), 1-17.

Coleman, R. D., Rapport, L. J., Millis, S. R., Ricker, J. H., & Farchione, T. J. (1998). Effects of coaching on detection of malingering on the California Verbal Learning Test. *Journal of Clinical and Experimental Neuropsychology, 20*(2), 201-210.

Curtiss, G., Vanderploeg, R. D., Spencer, J., & Salazar, A. M. (2001). Patterns of verbal learning and memory in traumatic brain injury. *Journal of the International Neuropsychological Society, 7*(5), 574-585.

De Guise, E., Leblanc, J., Feyz, M., & Lamoureux, J. (2005). Prediction of the level of cognitive functional independence in acute care following traumatic brain injury. *Brain Injury, 19*(13), 1087-1093.

Delis, D., Kramer, J., Kaplan, E., & Ober, B. (2000). The California Verbal Learning Test (2nd ed.): Psychological Assessment Resources, Inc.

DenBoer, J. W., & Hall, S. (2007). Neuropsychological test performance of successful brain injury simulators. *Clinical Neuropsychologist, 21*(6), 943-955.

Dikmen, S. S., Machamer, J. E., Winn, H. R., & Temkin, N. R. (1995). Neuropsychological outcome at 1-year post head-injury. *Neuropsychology, 9*(1), 80-90.

Draper, K., & Ponsford, J. (2008). Cognitive functioning ten years following traumatic brain injury and rehabilitation. *Neuropsychology, 22*(5), 618-625.

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.

Flaro, L., Green, P., & Robertson, E. (2007). Word Memory Test failure 23 times higher in mild brain injury than in parents seeking custody: The power of external incentives. *Brain Injury, 21*(4), 373-383.

Fleming, J., Riley, L., Gill, H., Gullo, M. J., Strong, J., & Shum, D. (2008). Predictors of prospective memory in adults with traumatic brain injury. *J Int Neuropsychol Soc, 14*(5), 823-831.

Forrester, G., Encel, J., & Geffen, G. (1994). Measuring posttraumatic amnesia (pta) - an historical review. *Brain Injury, 8*(2), 175-184.

Gierok, S. D., & Dickson, A. L. (2000). TOMM: Test of memory malingering. [Book Review]. *Archives of Clinical Neuropsychology, 15*(7), 649-651.

Green, P. (2005). Medical Symptom Validity Test. Edmonton: Green's Publishing.

Green, P. (2008). Non-verbal Medical Symptom Validity Test. Edmonton: Green's Publishing.

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

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

Greiffenstein, M. F., Gola, T., & Baker, W. J. (1995). MMPI-2 validity scales versus domain-specific measures in detection of factitious traumatic brain injury. *Clinical Neuropsychologist, 9*(3), 230-240.

Greve, K. W., Binder, L. M., & Bianchini, K. J. (2009). Rates of Below-Chance Performance in Forced-Choice Symptom Validity Tests. *Clinical Neuropsychologist, 23*(3), 534-544.

Greve, K. W., Ord, J., Curtis, K. L., Bianchini, K. J., & Brennan, A. (2008). *Detecting malingering in traumatic brain injury and chronic pain: A comparison of three forced-choice symptom validity tests*.

Gronwall, D., & Wrightson, P. (1981). Memory and information processing capacity after closed head injury. *J Neurol Neurosurg Psychiatry, 44*(10), 889-895.

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized Linear Models and Extensions* (2nd ed.). College Station, TX: Stata Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.): Springer.

Heaton, R. K., Smith, H. H., Jr., Lehman, R. A., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *J Consult Clin Psychol, 46*(5), 892-900.

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

Himanen, L., Portin, R., Isoniemi, H., Helenius, H., Kurki, T., & Tenovuo, O. (2006). Longitudinal cognitive changes in traumatic brain injury - A 30-year follow-up study. *Neurology, 66*(2), 187-192.

Hiscock, M., & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11*(6), 967-974.

Hoofien, D., Gilboa, A., Vakil, E., & Donovick, P. J. (2001). Traumatic brain injury (TBI) 10-20 years later: a comprehensive outcome study of psychiatric symptomatology, cognitive abilities and psychosocial functioning. *Brain Injury, 15*(3), 189-209.

Horner, M. D., Bedwell, J. S., & Duong, A. (2006). Abbreviated form of the Test of Memory Malingering. *International Journal of Neuroscience, 116*(10), 1181-1186.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logisitic Regression*. New York, NY: Wiley Interscience.

Iverson, G. L. (1995). Qualitative aspects of malingered memory deficits. *Brain Injury, 9*(1), 35-40.

Iverson, G. L. (2006). Ethical issues associated with the assessment of exaggeration, poor effort, and malingering. *Applied Neuropsychology, 13*(2), 77-90.

Iverson, G. L., Slick, D. J., & Franzen, M. D. (2000). Evaluation of a WMS-R malingering index in a non-litigating clinical sample. *Journal of Clinical and Experimental Neuropsychology, 22*(2), 191-197.

Kashluba, S., Hanks, R. A., Casey, J. E., & Millis, S. R. (2008). Neuropsychologic and functional outcome after complicated mild traumatic brain injury. *Arch Phys Med Rehabil, 89*(5), 904-911.

Langeluddecke, P. M., & Lucas, S. K. (2003). Quantitative measures of memory malingering on the Wechsler Memory Scale - Third edition in mild head injury litigants. *Archives of Clinical Neuropsychology, 18*(2), 181-197.

Langlois, J. A., Rutland-Brown, W., & Thomas, K. E. (2006). Traumatic Brain Injury in the United States: Emergency Department Visits, Hospitalizations, and Deaths. Atlanta, GA: Center for Disease Control and Prevention.

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *Clinical Neuropsychologist, 17*(3), 410-425.

Larrabee, G. J. (2005). Mild Traumatic Brain Injury. In G. J. Larrabee (Ed.), *Forensic Neuropsychology: A scientfic approach*. New York, NY: Oxford University Press.

Larrabee, G. J. (2007). Introduction: Malingering, Research Designs, and Base Rates. In G. J. Larrabee (Ed.), *Assessment of Malingered Neurocognitive Deficits* (pp. 3 - 13). New York: Oxford University Press.

Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *Clinical Neuropsychologist, 22*(4), 666-679.

Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2009). 40 Plus or Minus 10, a New Magical Number: Reply to Russell. *Clinical Neuropsychologist, 23*(5), 841-849.

Levin, H. S. (1992). Neurobehavioral recovery. *Journal of Neurotrauma, 9*, S359-S373.

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, J., & Fischer, J. S. (2004). *Neuropsychological Assessment* (4th ed.): Oxford University Press.

Majer, M., Welberg, L. A. M., Capuron, L., Miller, A. H., Pagnoni, G., & Reeves, W. C. (2008). Neuropsychological performance in persons with chronic fatigue syndrome: Results from a population-based study. *Psychosomatic Medicine, 70*(7), 829-836.

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.

McAllister, T. W. (1992). Neuropsychiatric sequelae of head injuries. [; Review]. *Psychiatr Clin North Am, 15*(2), 395-413.

McKinlay, W. W., Brooks, D. N., Bond, M. R., Martinage, D. P., & Marshall, M. M. (1981). The short-term outcome of severe blunt head injury as reported by relatives of the injured persons. *J Neurol Neurosurg Psychiatry, 44*(6), 527-533.

Merten, T., Green, P., Henry, M., Blaskewitz, N., & Brockhaus, R. (2005). *Analog validation of German-language symptom validity tests and the influence of coaching*.

Miller, J. B., Millis, S. R., & Fichtenberg, N. L. (2010). Diagnostic Efficiency of an Ability Focused Assessment Battery with a mixed clinical sample. *Clinical Neuropsychologist, 24*(4), 678-688.

Millis, S. R. (1994). Assessment of motivation and memory with the recognition memory test after financially compensable mild head-injury. *Journal of Clinical Psychology, 50*(4), 601-605.

Millis, S. R., Putnam, S. H., Adams, K. M., & Ricker, J. H. (1995). The California Verbal Learning Test in the detection of incomplete effort in neuropsychological evaluation. *Psychological Assessment, 7*(4), 463-471.

Millis, S. R., & Ricker, J. H. (1994). Verbal-learning patterns in moderate and severe traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 16*(4), 498-507.

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*(8), 1094-1102.

Nakase-Richardson, R., Yablon, S. A., & Sherer, M. (2007). Prospective comparison of acute confusion severity with duration of post-traumatic amnesia in predicting employment outcome after traumatic brain injury. *Journal of Neurology Neurosurgery and Psychiatry, 78*(8), 872-876.

Nelson, N. W., Boone, K., Dueck, A., Wagener, L., Lu, P., & Grills, C. (2003). Relationships between eight measures of suspect effort. *Clinical Neuropsychologist, 17*(2), 263-272.

Ord, J. S., Greve, K. W., & Bianchini, K. J. (2008). Using the Wechsler Memory Scale-III to detect malingering in mild traumatic brain injury. *Clin Neuropsychol, 22*(4), 689-704.

Parker, S. A., & Serrats, A. F. (1976). Memory recovery after traumatic coma. *Acta Neurochir (Wien), 34*(1-4), 71-77.

Pearson Education. (2008). Wechsler Memory Scale, Fourth Edition: Clinical features of the new edition.

Pereira, A. P. A. (2007). Assessment of memory in rehabilitation counseling. *Journal of Rehabilitation, 73*(2), 15-25.

Piantadosi, S. (2005). *Clinical Trials, A Methodological Perspective* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*(1), 33-65.

Raftery, A. E. (1996). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological Methodology* (Vol. 25, pp. 111-163). Oxford, UK: Basil Blackwell.

Rapport, L. J., Axelrod, B. N., & Mansharamani, N. H. (1995). Self-generated errors on Wechsler Memory Scale Associate Learning. *Clinical Neuropsychologist, 9*, 278.

Rapport, L. J., Axelrod, B. N., Theisen, M. E., Brines, D. B., Kalechstein, A. D., & Ricker, J. H. (1997). Relationship of IQ to verbal learning and memory: Test and retest. *Journal of Clinical and Experimental Neuropsychology, 19*(5), 655-666.

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

Richman, J., Green, P., Gervais, R., Flaro, L., Merten, T., Brockhaus, R., & Ranks, D. (2006). Objective tests of symptom exaggeration in independent medical examinations. *Journal of Occupational and Environmental Medicine, 48*(3), 303-311.

Roche, N. L., Moody, A., Szabo, K., Fleming, J. M., & Shum, D. H. K. (2007). Prospective memory in adults with traumatic brain injury: An analysis of perceived reasons for remembering and forgetting. *Neuropsychological Rehabilitation, 17*(3), 314-334.

Rogers, R. (1990). Models of feigned mental-illness. *Professional Psychology-Research and Practice, 21*(3), 182-188.

Rogers, R. (1997a). Introduction. In R. Rogers (Ed.), *Clinical Assessment of Memory and Malingering* (2nd ed.). New York, NY: The Guilford Press.

Rogers, R. (1997b). Researching Dissimulation. In R. Rogers (Ed.), *Clinical Assessment of Malingering and Deception* (2nd ed., pp. 398 - 426). New York, NY: The Guilford Press.

Rohling, M. L., Meyers, J. E., & Millis, S. R. (2003). Neuropsychological impairment following traumatic brain injury: A dose-response analysis. *Clinical Neuropsychologist, 17*(3), 289-302.

Ruff, R. M., & Richardson, A. M. (1999). Mild Traumatic Brain Injury. In J. J. Sweet (Ed.), *Forensic Neuropsychology: Fundamentals and practice*. Lisse, Netherlands: Swets and Zeitlinger.

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), 294-299.

Russeler, J., Brett, A., Klaue, U., Sailer, M., & Munte, T. F. (2008). The effect of coaching on the simulated malingering of memory impairment. *Bmc Neurology, 8*, 14.

Sherer, M., & Madison, C. F. (2005). Moderate and Severe Traumatic Brain Injury. In G. J. Larrabee (Ed.), *Forensic Neuropsychology: A scientfic approach*. New York, NY: Oxford University Press.

Shores, E. A., Lammel, A., Hullick, C., Sheedy, J., Flynn, M., Levick, W., & Batchelor, J. (2008). The diagnostic accuracy of the Revised Westmead PTA Scale as an adjunct to the Glasgow Coma Scale in the early identification of cognitive impairment in patients with mild traumatic brain injury. *Journal of Neurology Neurosurgery and Psychiatry, 79*(10), 1100-1106.

Silverberg, N. D., Wertheimer, J. C., & Fichtenberg, N. L. (2007). An effort index for the repeatable battery for the assessment of neuropsychological status (RBANS). *Clinical Neuropsychologist, 21*(5), 841-854.

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *Clinical Neuropsychologist, 13*(4), 545-561.

Stevens, A., Friedel, E., Mehren, G., & Merten, T. (2008). Malingering and uncooperativeness in psychiatric and psychological assessment: Prevalence and effects in a German sample of claimants. *Psychiatry Research, 157*(1-3), 191-200.

Stringer, A. Y., & Nadolne, M. J. (2000). Neuropsychological assessment: Contexts for contemporary clinical practice. In G. Groth-Marnat (Ed.), *Neuropsychological assessment in clinical practice: A guide to test interpretation and integration*. New York: Wiley.

Suhr, J. A., & Barrash, J. (2007). Performance on Standard Attention, Memory, and Psychomotor Speed Tasks as Indicators of Malingering. In G. J. Larrabee (Ed.), *Assessment of Malingered Neurocognitive Deficits* (pp. 131 - 170). New York: Oxford University Press.

Suhr, J. A., & Gunstad, J. (2007). Coaching and Malingering: A Review. In G. J. Larrabee (Ed.), *Assessment of Malingered Neurocognitive Deficits* (pp. 287 - 311). New York: Oxford University Press.

Tabaddor, K., Mattis, S., & Zazula, T. (1984). Cognitive sequelae and recovery course after moderate and severe head injury. *Neurosurgery, 14*(6), 701-708.

Tate, D., Kalpakjian, C., & Kwon, C. (2008). The use of randomized clinical trials in rehabilitation psychology. *Rehabilitation Psychology, 53*(3), 268-278.

Teichner, G., Wagner, M. T., & Newman, S. A. (2000). Psychometric validation and clinical application of the Test of Memory Malingering (TOMM). *Archives of Clinical Neuropsychology, 15*(8), 673-674.

Thorburm, W. M. (1915). Occam's Razor. *Mind, 24*, 287 - 288.

Tombaugh, T. (1996). *Test of Memory Malingering*. North Tonawanda, NY: Multi-Health Systems.

Tombaugh, T. (1997). The test of memory malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment, 9*(3), 260-268.

Vakil, E. (2005). The effect of moderate to severe traumatic brain injury (TBI) on different aspects of memory: a selective review. *J Clin Exp Neuropsychol, 27*(8), 977-1021.

van Gorp, W. G., Humphrey, L. A., Kalechstein, A., Brumm, V. L., McMullen, W. J., Stoddard, M., & Pachana, N. A. (1999). How well do standard clinical neuropsychological tests identify malingering? A preliminary analysis. *Journal of Clinical and Experimental Neuropsychology, 21*(2), 245-250.

Vanderploeg, R. D., Crowell, T. A., & Curtiss, G. (2001). Verbal learning and memory deficits in traumatic brain injury: Encoding, consolidation, and retrieval. *Journal of Clinical and Experimental Neuropsychology, 23*(2), 185-195.

Warrington, E. K. (1984). Recognition Memory Test. Lutz, FL: Psychological Assessment Resources.

Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology, 19*, 87 - 95.

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

Williams, J. M. (1998). The malingering of memory disorder. In C. R. Reynolds (Ed.), *Detection of malingering during head-injury litigation*. New York: Plenum.

Wolfe, P. L., Millis, S. R., Hanks, R., Fichtenberg, N., Larrabee, G. J., & Sweet, J. J. (2010). Effort Indicators within the California Verbal Learning Test-II (CVLT-II). *Clinical Neuropsychologist, 24*(1), 153-168.

ABSTRACT

ASSESSMENT OF MEMORY FUNCTION AND EFFORT USING THE WECHSLER MEMORY SCALE – 4TH EDITION

by

JUSTIN MILLER, M.A.

May 2011

Advisor: Dr. R. Douglas Whitman

Major: Psychology (Clinical)

Degree: Doctor of Philosophy

Even the most psychometrically sound measures are sensitive to the level of effort put forth by the examinee and their intent. This is especially true for measures of memory functioning that are a common target of negative response bias and withholding effort. The aim of the present study was to develop methods for detecting these behaviors for the current edition of the Wechsler Memory Scale, 4th Edition (WMS-IV) using a community sample of healthy adults coached to simulate traumatic brain injury (TBI) and a sample of bona fide TBI survivors. The primary analytic strategy involved generation of prediction models to classify participants according to group membership via logistic regression and evaluate classification accuracy with receiver operating characteristic (ROC) curves. It was predicted that measures from within the WMS-IV would be able to reliably differentiate between actual and simulated TBI, and furthermore, between well simulated and poorly simulated TBI. The results from this study provide confirming evidence in support of both tested hypotheses. Several key findings can be taken from the present study. First there is strong support for the use of multiple measures of symptom validity interspersed throughout testing. Second, this

study provides firm support for the use of the WMS-IV in detecting cases of feigned cognitive impairment and offered several methods to effectively do so. Third, the present study has also shown just how difficult it is to simulate a TBI in a way that is credible. Fourth, these results highlight the similarity of performance profiles between poor effort and intentional negative response bias. Several areas of future research are presented.

AUTOBIOGRAPHICAL STATEMENT

Education

May 2007	Master of Arts Wayne State University, Detroit, Michigan Thesis: Serial semantic priming: Exploring the role of inhibition between related semantic networks Faculty Advisor: R. Douglas Whitman, Ph.D.
April 2003	Bachelor of Arts University of Michigan, Dearborn, Michigan Major: Psychology

Clinical Experience

July 2010 – June 2011	Semel Institute for Neuroscience and Human Behavior University of California, Los Angeles Pre-doctoral Internship, Clinical Neuropsychology Advisor: Robert Bilder, Ph.D.
August 2004 – June 2010	Wayne State University Psychology Clinic Individual Psychological Assessments; Individual & Group psychotherapy
September 2007 – August 2009	Rehabilitation Institute of Michigan Inpatient & Outpatient Neuropsychological Assessment and individual psychotherapy

Research Experience

September 2009 – June 2010	Research Assistant/Psychometrist Southeastern Michigan Traumatic Brain Injury Model System (SEMTBIS) Robin Hanks, Ph.D., ABPP-ABCN Chief of Rehabilitation Psychology and Neuropsychology Detroit, Michigan
August 2004 – August 2007	Research Assistant R. Douglas Whitman, Ph.D., Professor Psychology Departmental Chair, Wayne State University Detroit, Michigan
May 2002 – July 2004	Research Assistant Bruno Giordani, Ph.D., Associate Professor, Neuropsychology Division Director; University of Michigan Health System, Ann Arbor, Michigan

Competitive Funding:

2009 - 2010	Principal Investigator: Assessment of Memory Function and Effort using the Wechsler Memory Scale – 4th Ed. Del Harder Foundation for rehabilitation research, Detroit Medical Center ($19,915)
2008 – 2009	Graduate-Professional Competitive Scholarship - Wayne State University
2004 – 2005	Thomas C. Rumble Fellowship – Wayne State University