1-1-2012

Detecting suboptimal effort in traumatic brain injury assessment

Jesse Ryan Bashem
Wayne State University,

DETECTING SUBOPTIMAL EFFORT IN TRAUMATIC BRAIN INJURY ASSESSMENT

by

JESSE BASHEM

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF ARTS

2011

MAJOR: PSYCHOLOGY

Approved by:

_____

Advisor Date

ACKNOWLEDGEMENTS

Lisa J. Rapport, Ph.D.

Scott Millis, Ph.D.

Robin Hanks, Ph.D.

Justin Miller, Ph.D.

Bradley N. Axelrod, Ph.D.

R. Douglas Whitman, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Recent estimates of the base rate for malingering in forensic cases involving post-concussive neurocognitive deficits, such as memory impairment associated with mild head injury, approach 40%. As the role of neuropsychological assessment increases in medico-legal referral contexts, the demand for research evaluating performance effort greatly increases. Hence, the primary goal of the proposed research is to enhance diagnostic accuracy regarding identification of bona fide traumatic brain injury (TBI) versus feigned neurocognitive impairment. Purposeful presentation of suboptimal effort is a primary pitfall to accurate assessment, especially among individuals seeking compensation. It is known that successful simulation of deficits becomes increasingly difficult when feigning is required across multiple measures. This logic lays the foundation for the "patterns of performance" theory: A multidimensional, multi-method approach will likely increase detection rates as compared to the interpretation of isolated test scores. However, studies comparing concordance of multiple indices to assess effort in TBI assessment are sparse, as are studies employing an external criterion (i.e., "known-groups" designs). This lack of knowledge is an important problem because patients who provide insufficient effort (e.g., feign impairment) unfairly stress the legal and healthcare systems, whereas patients who are inaccurately labeled as malingerers are unjustly accused as fakers and unfairly denied the resources and services they deserve.

The central hypothesis was that comparing multiple measures of performance effort, used in various combinations, would yield a test battery that provides the most clinically efficient and valid classification accuracy of bona fide and feigned cognitive impairment. The hypothesis was formulated on preliminary data collected for this study, as well as on psychometric and statistical

theory regarding detection, and clinical research on malingering. The hypothesis was tested using a known-groups design that incorporates 57 adults with bona fide TBI and 60 healthy adult simulators. The study compared the clinical utility and classification accuracies of six symptom validity tests (SVTs) designed to assess for suboptimal effort among examinees who participate in a standardized assessment of memory (see Research Design).

Specific Aim 1: Identify the diagnostic validity and classification accuracy statistics for each SVT in isolation. The hit rate, sensitivity, and specificity for each SVT were calculated. The working hypothesis was that each of the SVTs would show at least moderate sensitivity to the presence of insufficient effort in the context of detecting feigned memory impairment from TBI.

Specific Aim 2: Compare the incremental, clinical utility of each SVT in relation to one another. Concordance rates between classifications made with each SVT were compared, and the interrelationships among the measures were determined.

Specific Aim 3: Determine combinations of SVTs that produce the most efficient, diagnostically valid index of suboptimal effort. The working hypothesis was that incremental validity is enhanced with the inclusion of measures using multiple, distinct methods for assessing a similar trait. Binary logistic regression, receiver operating characteristic (ROC) curves, and Bayesian information criterion (BIC) statistics were employed to determine the best-fit model.

This study investigated empirically-supported combinations of indices resulting in the greatest classification accuracy. The topic has important implications for clinical assessment in compensation-seeking contexts, especially as it may maximize valid allocation of healthcare and other resources to persons with bona fide TBI.

Background and Significance

With approximately 1.7 million new injuries each year and 5.3 million people living with

injury-related deficits, traumatic brain injury (TBI) is a significant health problem in the United States (Faul et al., 2010; Finelstein, Corso, & Miller, 2006). TBI can result in an array of complex, variable, and long-lasting cognitive deficits. Memory impairments are especially common and long-lasting following TBI (Lezak, Howieson, Loring, & Hannay, 2004). Although formal cognitive evaluations routinely include standardized measures of memory (Lezak et al., 2004), the validities of these tools are vulnerable to the level of effort provided by the examinee during testing. Suboptimal effort may result for a myriad of reasons, both conscious and unconscious (Lynch, 2004). Without accurate means of assessing effort, clinicians are left with test results of questionable validity. Invalid assessments can then lead to a wide assortment of negative medical and legal consequences, including misdiagnoses, improper intervention strategies, inaccurate outcomes from treatment efficacy studies, and unfair allocation of resources and monetary compensation. An array of empirically-derived tests and detection strategies has been generated to assess for suboptimal effort. Unfortunately, little research has examined consolidating these methods to derive the most clinically efficient and statistically powerful prediction models.

Clinical Need

Rehabilitation healthcare professionals and patients collaborate in service to develop strategies to overcome the functional limitations of TBI and maximize preserved abilities. The capacity to make successful recommendations, however, is contingent on accurate measurement of the functional abilities of the patient; hence, there exists a need to supply state-of-the-science assessments, referrals, and rehabilitation services to TBI survivors. Concurrently, increased public awareness of cognitive deficits following even mild TBI has given rise to an increasing

number of individuals seeking medico-legal compensation for damages (Pankratz & Binder, 1997). TBI-related cases account for a majority of all neuropsychological forensic cases. Strikingly, current estimates approximate 30% of civil cases, 20% of criminal cases, and 10% of medical cases as suspect of suboptimal effort or feigned impairments, memory deficits being the most commonly reported (Mittenberg, Patton, Canyock, & Condit, 2002). Developed specifically for use in neuropsychological contexts, the most widely-used diagnostic scheme includes the criteria for malingered neurocognitive deficit (MND) proffered by Slick, Sherman and Iverson (1999); these authors 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" (p. 552). Base rates differ across clinical settings; however, it has been estimated that 30-40% of mild TBI cases in which compensation is sought are likely malingering the impairment (Larrabee, 2003; Binder & Kelly, 1994). As stipulated in the official position of the National Academy of Neuropsychology (Bush et al., 2005) and the American Academy of Clinical Neuropsychology (Heilbronner, Sweet, Morgan, Larrabee, & Millis, 2009), assessment of symptom validity is an essential aspect of all neuropsychological evaluations and demands greater attention by researchers.

Assessment of Effort

The accuracy of interpretations made from test data that purportedly reflect cognitive abilities relies fundamentally on the assumption that examinees have responded with sincere and adequate effort. Especially for persons who sustain mild TBI, the results of neuropsychological tests are frequently the sole source of objective evidence that brain injury has occurred because neurologic exams and neuroimaging data are often negative (Constantinou, Bauer, Ashendorf,

Fisher, & McCaffrey, 2005; Lynch, 2004). Presently, a large number of stand-alone symptom validity tests (SVT) are commonly used during neuropsychological evaluations. Of the measures specifically designed to assess for suboptimal effort, the symptom validity test (SVT) paradigm is the most popular among neuropsychologists (Constantinou et al., 2005; Slick, Hopp, Strauss, & Spellacy, 1996). Although published by independent parties, these tests share two common features: (1) they are related to aspects of memory performance as this cognitive domain is highly susceptible to impression management among persons undergoing neuropsychological evaluation for TBI; and (2) they employ a two-alternative forced-choice format that utilizes the known probabilities of correct responding given no prior exposure to the test stimuli (Hiscock & Hiscock, 1989). Across the individual tests, this type of SVT is interpreted using empirically-derived cutoff scores that generally yield good specificity and positive predictive value, yet tend to be limited in sensitivity (Binder & Kelly, 1994). The Test of Memory Malingering (TOMM) (Tombaugh, 1996) and the Medical Symptom Validity Test (MSVT) (Green, 2005) are examples of the two-alternative forced-choice type of SVT that are widely employed by clinicians and researchers (Sharland & Gfeller, 2007; Slick et al., 2004; Richman et al., 2006).

Initial validation studies of the MSVT, TOMM, and similar SVTs indicate that they are robust to the effects of age, education, TBI, dementia, depression, and anxiety (Constantinou et al., 2005; Green, 2005; Rees et al., 2001); also, research demonstrates that the SVT paradigm may be resilient to psychosis as well (Schroeder & Marshall, 2011). Use of an SVT specifically is recommended as standard practice in a neuropsychological assessment (Inman & Berry, 2002; Binder & Kelly, 1994). At least one study suggests that poor performance on this type of recognition memory SVT was sensitive to generalized poor effort on cognitive domains other than memory (Heilbronner et al., 2009). However, research examining this issue of construct

generalizability is sparse and can suffer from contamination of criterion with the predictor when classification of the groups is based on SVT performance itself. Furthermore, many stand-alone measures are highly susceptible to coaching and can be easily identified by examinees as measures of effort or malingering. For example, during debriefing, fewer than 10% of examinees instructed to feign TBI rated the TOMM as a measure of ability, recognizing it as a measure of effort, instead (Tan et al., 2002). In response to the latter problem, several "embedded" measures of effort have also been identified: these are indices derived from standard ability tests commonly administered in a neuropsychological battery (i.e., "built in") that signify non-credible or "suspect" performance. Because they are embedded, they may be less obvious and less susceptible to coaching than stand-alone measures. Indices based on the Digit Span subtest of the (now outdated) WAIS-III have shown to be helpful in detecting suboptimal effort, likely reflecting that extremely poor performance on Digit Span is relatively rare among non-litigating TBI patients with various types of brain damage and even significant impairments (Bush et al., 2005). The California Verbal Learning Test–2nd Edition (CVLT-2) Forced-Choice Recall trial also has received focus as modestly capable of detecting suboptimal effort while likely going undetected as such by examinees (Wolfe, Millis, Hanks, Fichtenberg, Larrabee, & Sweet, 2010). The consensus of the flagship organizations and experts in the field is that both stand-alone effort tests and embedded validity indicators should be used (Bush et al., 2005).

Classification of persons as malingerers is based on a variety of different methods, each of which is prone to measurement error. Based on the patterns of performance theory, a multidimensional, multi-method approach to detecting malingering of neuropsychological deficits will likely increase detection rates of malingerers as compared to the interpretation of isolated test scores. Models that combine multiple indices are an especially promising method of

enhancing classification accuracy (Richman et al., 2006; Larrabee, 2008). Historically, individual effort indices have been selected that yield high specificity at the expense of relatively low sensitivity (Bush et al., 2005). In this clinically applied context, specificity and negative predictive power outweigh sensitivity and positive predictive power, and the rationale for this position is quite evident; the worst-case scenario is to deny resources and compensation to someone who should have rightful access to them, especially in the stigmatizing context of falsely labeling the person as purposefully dishonest. Within the domain of TBI rehabilitation, withholding treatment from a person in need is a more severe offense than supplying treatment or other resources to an individual feigning impairment. However, given limited healthcare resources and the rise in forensic cases seeking monetary and other compensation, there exists a strong need to increase a test battery's sensitivity, thus allowing clinicians to identify feigned or grossly exaggerated impairments accurately. Psychometrically, combining the results of multiple indices (i.e., chaining likelihood ratios) will provide increased sensitivity as compared to single indicators without detracting from specificity, providing that the individual measures are independent or have low intercorrelations (Rees et al., 2001).

Limitations of the extant literature

Despite establishing the standard of prevailing practice to include multiple indices of effort (Binder & Kelly, 1994; Bush et al., 2005; Slick, Sherman, & Iverson, 1999), studies comparing concordance of multiple indices are sparse, and fewer still have attempted to combine multiple indices into a single predictive formula (Greiffenstein, Greve, Bianchini, & Baker, 2008). Those existing studies largely report on now-outdated measures deemed unacceptable for modern practice such as the WAIS-III (now WAIS-IV), the CVLT (now CVLT-2), and the Rey

15-Item Test, which has been repeatedly shown to be ineffective (Vallabhajosula & van Gorp, 2001). As a result, calls for research aimed at the diagnostic efficiency of incorporating multiple measures of effort have resounded throughout the neuropsychological community.

Studies employing an external criterion (i.e., "known-groups" designs) also are relatively sparse. Much research in malingering employs “analog design”: a simulation paradigm in which healthy adults assigned to feign TBI (and sometimes coached in how to succeed) are compared to healthy adults instructed to put forth best effort. Among the many strengths of this design is the level of experimental control; however, the design is faulted for having relatively low ecological validity as compared to known-group designs (Rogers, 1988). Studies that include groups of persons with bona fide TBI are far fewer than analog studies; they tend to report on relatively small samples with a limited selection of effort measures (Greve, Ord, Curtis, Bianchini, & Brennan, 2008). Very few studies have employed these paradigms simultaneously so that effort measures could be examined together in this context. Also surprisingly absent are explorations regarding characteristics of misclassified cases in the bona fide TBI group, as are explorations of effort measures as a continuous phenomenon versus dichotomous classification and how these relate to performances in other domains among bona fide TBI and examinees known to demonstrate low effort.

CHAPTER 2

METHOD

Participants

Participants were 57 adults with TBI, recruited from the pool of participants enrolled in the Southeastern Michigan Traumatic Brain Injury System (SEMTBIS), which is part of the TBI Model Systems (TBIMS) program funded by the National Institute on Disability and Rehabilitation Research. Inclusion criteria for the SEMTBIS research project stipulates that all participants have incurred a moderate to severe TBI as indicated by the following: post-traumatic amnesia lasting at least 24 hours, loss of consciousness for at least 30 minutes, Glasgow Coma Scale score less than 13 upon arrival to the emergency department, or the detection of abnormal intracranial status via neuroimaging. Further, participants must have received acute care within 72 hours of injury, been transferred to a rehabilitation unit, and have been at least 16 years old at the time of injury. As a result of inclusion criteria, the sample excludes persons with mild injuries or very severe brain injuries who did not receive inpatient rehabilitation. SEMTBIS participants who agreed to be contacted for future research projects were notified of an opportunity to participate in the current study by the SEMTBIS research coordinator. Interested individuals were screened for eligibility and scheduled by the principle investigator.

A demographically-comparable sample of neurologically healthy adults (n = 60) were recruited for the TBI simulator group from the Southeastern Michigan area. Recruitment was conducted via newspaper advertisements, online postings, and flyer postings throughout the Wayne State University campus. Exclusion criteria included history of neurologic illness or injury (e.g., TBI, concussion, stroke, or seizure disorder).

Complete demographic data for each group and the total sample are presented in Table 1:

also included are the descriptive statistics for the effort indices. The TBI group (n = 57) was predominantly African-American (82.5%) men (93.0%) with a mean age of 44.6 years (SD = 11.9) and mean education of 12.2 years (SD = 2.1). As predicted by the WTAR, mean estimated IQ for the group was 85.2 (SD = 9.1). The SIM group (n = 60) also was primarily African-American (68.3%) men (86.7%), with a mean age of 44.0 years (SD = 11.4), mean education of 12.7 years (SD = 2.0), and an estimated IQ of 92.9 (SD = 12.5). To avoid a significant difference between the groups in years of education (TBI < SIM), cases with fewer than 9 years of education were excluded and the present sample was limited in range of education from 9 – 21 years. Comparisons of the groups found no significant differences on age (F[1,116] = 0.03, p = .79), education (F[1,116] = 2.14, p = .15), or proportion of men (X^2[1, N = 117] = 1.27, p = .26).

Descriptive correlations between demographic variables and the effort indices are presented for the Simulator Group (Table 2a), TBI Group (Table 2b) and Total Sample (Table 2c): age, education, estimated IQ, and (for TBI participants) injury severity (Glasgow Coma Scale at admission to the ER) and months since injury. Among the Simulator group (Table 2a), there were no significant correlations to age, years of education, or estimated IQ. Among the TBI Group are modest correlations between Reliable Digit Span and both estimated IQ (rho = .44) and months since injury (rho = -.36), and between the total number of effort measures failed and both age (rho = .30) and months since injury (rho = .36).

Measures

Injury Severity. The motor subscale of the Glasgow Coma Scale (GCS) was employed as a measure of TBI severity. Specifically, brain injury severity was represented by the time required to follow commands, as indicated by the number of days needed to twice obtain a score of 6 on the GCS motor subscale within a 24-hour period (Dikmen, Machamer, Winn, & Temkin,

1995; Rohling, Meyers, & Millis, 2003).

Premorbid Intelligence. The Wechsler Test of Adult Reading (WTAR) (The Psychological Corporation, 2001) is a word reading test that consists of 50 irregular words to pronounce aloud. Recognition reading vocabulary is relatively robust to neurologic impairment and has been shown to be an excellent estimate of overall intellectual ability, or Full Scale IQ (Johnstone, Hexum, & Ashkanazi, 1995). Past research has used the WTAR to generate estimates of intellectual functioning among people with TBI (Green, Melo, Christensen, Ngo, Monette, & Bradbury, 2008).

Effort: Memory Specific Symptom Validity Measures

Test of Memory Malingering (TOMM) (Tombaugh, 1996). This 50-item, forced-choice measure uses visual recognition of drawings to assess an examinee's level of effort and is commonly used in psychological assessment batteries. The test consists of two learning trials, both of which present the same 50, hand-drawn stimulus items in different orders. Each trial is followed by a forced-choice task that presents a previously shown item alongside a novel foil item, and the patient is asked to choose the item they remember having seen previously. An optional retention trial is also included following the prior two trials. Totaling the correct responses in each trial derives two continuous scores that can be compared to statistically-derived (below chance) cut scores for each trial. Typically, effort research relies on examining the performance on Trial 2, with an obtained score of less than 45 signifying inadequate effort. Although the TOMM has shown adequate specificity in detecting suboptimal effort (Gierok & Dickson, 2000; Rees, Tombaugh, Gansler, & Moczynski, 1998; Teichner, Wagner, & Newman, 2000), research also indicates that the level of sensitivity it provides may be too low to use alone (Greve et al., 2008).

Medical Symptom Validity Test (MSVT) (Green, 2005). An abbreviated version of the Word Memory Test, the MSVT is a computerized measure of verbal memory that utilizes 10 word pairs across four subtests (e.g. immediate recognition, delayed recognition, paired association, and free recall) to assess memory as well as the examinee's overall response consistency. This test also is commonly used in psychological assessment batteries. Validation studies of the test's embedded measures support the ability of the MSVT to detect suboptimal effort (Richman et al., 2006; Merten, Green, Henry, Blaskewitz, & Brockhaus, 2005). The test consists of 10 word pairs that are presented twice over the course of 5 minutes, immediately followed by three response trials (a free-recall trial [IR], a paired-associate recall trial, and an immediate forced-choice recognition trial) that vary in difficulty from "easy" to "hard." After a 10-minute delay, a forced-choice trial (DR) is administered. These four scores are complemented by a fifth "consistency score" (CNS) used to assess the stability of performance between the aforementioned trials. The MSVT yields five scores, three of which are used to assess for adequate effort: Immediate Recall (IR), Delayed Recall (DR), and Consistency (CNS). As published in the test manual, scores $\leq$ 85% on *any* of the IR, DR, or CNS conditions is considered a failure on the test (Green, 2005). Examinees are required to attend to the stimulus words as they are presented, and then asked to immediately recall as many as possible while the examiner enters their responses into the computer (with the screen turned away from the examinee). Next, the examiner reads the stem word from each of the pairs and asks the examinee to supply the paired word. Control of the computer is then given back to the examinee and requires that they use an external mouse to select the target words from novel foils. After the 10-minute delay, the examinee once more completes a forced-choice trial that incorporates a new set of novel foils.

The standardized scoring method developed by the test publishers ultimately provides a

single, dichotomous index of effort; yet, unlike the other measures administered in this study, this categorical decision is made by a applying a single cutoff score to three (as opposed to one) continuous subtest scores (e.g., IR, DR, and CNS). Although standardized scoring produces a categorical pass/fail variable that can function as a dichotomous predictor in logistic regression models, no formally established method exists regarding the synthesis of the MSVT's three continuous subtests into a single continuous predictor. As a result, deciding which score to employ as the continuous covariate in logistic regression models of the MSVT is met with uncertainty. An initial solution proposed that one of the three indices simply be selected as the sole representative of the MSVT in continuous models; however, arbitrarily excluding any particular index might unfairly represent the test's predicative ability. In order to provide the MSVT with an optimal opportunity to perform against the other tests, all three continuous indices were considered for use in testing the MSVT in the traditional logistic regression models. Examination of the descriptive correlations between the three MSVT indices (see Tables 2a and 2b) showed that the subtests are very highly intercorrelated (e.g., rho .86 to .91 among Simulators). Given these large intercorrelations and that the published psychometrics of the MSVT rely on all three indices to determine outcome, it was decided that the calculated average of the three indices be used to create a new continuous index named MSVTavg. Theoretically, increasing the number of similar items within an index consequently increases the reliability of the scale; therefore, using the continuous MSVTavg index as the MSVT's representative covariate in traditional logistic regression models was not only considered acceptable, it appears to favors the MSVT psychometrically as well. Therefore, it was determined that the dichotomous MSVT variable derived from standardized scoring would be used as the representative covariate in all logistic regression models testing the MSVT's categorical discriminability, and that the

MSVTavg variable be employed as the representative predictor in all models testing the MSVT's continuous discriminability.

Word Choice Test (WCT) (Pearson Education, 2008). This test is included in the Advanced Clinical Solutions (ACS) package available for the WAIS-IV (Wechsler, 2008) and WMS-IV (Pearson Education, 2008). It is a 50-item, forced-choice measure developed to parallel the Warrington Recognition Memory Test, which has been used successfully for many years to evaluate suboptimal effort and response bias across a variety of clinical and forensic settings (Millis, 1992; Millis & Putnam, 1994). Fifty word cards are presented to the examinee. The examinee is then instructed to state whether they associate the word as being "pleasant" or "unpleasant." Following this trial, the examinee is instructed to select the target words from a page consisting of all 50 targets and 50 paired foils. An obtained score of 44 or less signified inadequate effort.

California Verbal Learning Test – 2nd Edition (CVLT-II) (Delis, Kramer, Kaplan, & Ober, 2000). The CVLT-2, specifically its Forced-Choice Recall trial (CVLT-FC) is used as an embedded measure of effort. This list-learning task presents 16 words orally and requires examinees to recall the words over the course of five trials. Following the five learning trials, a distracter set is introduced and the examinee is administered a short-delay free recall trial. Another free recall trial is administered following a 20-minute delay to assess long-term retention. A final 10-minute delay proceeds a forced-choice recognition task in which the examinee must choose between a word from the original list and a novel foil. As delineated by the test publishers, performance on the CVLT-FC trial of the CVLT-2 is not related to memory

performance but was intentionally included as an embedded measure of examinee effort or suspected malingering. Per the CVLT-2 manual, a recognition score of 14 or less suggests suboptimal examinee effort.

Reliable Digit Span (RDS) (Weschler, 2008; Greiffenstein, Gola, & Baker, 1995). This embedded index was originally developed for the WAIS-III, Digit Span subtest; it is now included as an embedded index in the ACS package. Examinees are read strings of digits that must be recalled in either the same order (Digits Forward), backwards (Digits Backwards), or in sequence of lowest to highest digit (Digit Sequencing). The Reliable Digit Span is calculated by summing the longest span of digits correctly recalled on both trials of the Digits Forward and Digits Backward conditions. The published cutoff of 7 or less was shown to be moderately sensitive and adequately specific to suboptimal effort.

Procedure

Traumatic Brain Injury group (TBI). Participants currently enrolled in the SEMTBIS project were notified of this research opportunity via the SEMTBIS project coordinator. Informed consent procedures were completed per institutional review board guidelines. Those expressing interest had consented to be contacted by telephone by the primary investigator. They were informed of the opportunity to participate in a research project aimed at studying the use of a new psychological assessment test. Persons with TBI (n = 57) who agreed to participate were notified that the 3-hour evaluation would take place at the Rehabilitation Institute of Michigan's main campus, Novi campus, or the primary investigator's research lab on the Wayne State University campus. Testing was completed in a single session. The TBI participants were

instructed to put forth their full effort on all measures administered. Testing began by administering the WTAR. To avoid confounds due to test order effects, the WMS-IV and CVLT-II were counterbalanced with the administration of the effort measures (e.g., TOMM, MSVT, CVLT-FC, and the Digit Span subtest). All participants received $30 as compensation for their time.

Traumatic Brain Injury Simulator group (SIM). Participants in the SIM group (*n* = 60) were recruited from the Southeastern Michigan area via newspaper advertisements, online postings, and flyers posted throughout the Wayne State University campus and screened for eligibility via telephone. Informed consent procedures were completed with all SIM participants per institutional review board guidelines. In order to gain an accurate estimate of intellectual functioning for SIM participants, they were instructed to put forth full effort before being administered the WTAR. Upon completion of the WTAR, all SIM participants were told that the remainder of the assessment would focus on the ability of a new memory measure to assess the level of effort put forth during testing. Participants were then be presented with a scenario indicating his or her involvement in litigation following a motor vehicle accident that resulted in a TBI. The scenario was read from a script that has been used successfully in prior research on simulation with designs similar to that of this study (DenBoer & Hall, 2007; Tombaugh, 1997). Based on recommendations by Suhr and Gunstad (2007) regarding simulation research designs, all SIM participants were then provided with a pamphlet summarizing the nature of a TBI and the symptoms commonly associated with this type of injury such as slowed thinking, memory dysfunction, etc. (Coleman, Rapport, Millis, Ricker, & Farchione, 1998; Rapport, Farchione, Coleman, & Axelrod, 1998). Each participant was given as much time as needed to read over the material followed by an informal quiz to ensure adequate comprehension of the supplied

literature. Lastly, all SIM participants were informed that tests designed to measure effort will be included throughout the assessment. Administration of the test battery took place over the course of one 3-hour session; as with the TBI group, order of test administration was counterbalanced. All participants received $30 as compensation for their time.

Debriefing. Upon completion of the battery, all SIM participants were asked to complete a 6-item survey asking whether they intentionally tried to fake a TBI, how difficult it was to do so, and what strategies they used to simulate impairment. Data from any SIM participants who endorsed that they did not try to fake a TBI were excluded from analysis. The remainder of the responses will be catalogued by the examiner for use in future qualitative analyses of simulation strategies.

Statistical Analyses

Primary Analyses

Specific Aim 1. Examine the diagnostic validity and classification accuracy statistics for each SVT in isolation. The initial focus of Aim 1 centered on determining the classification accuracy statistics for each of the core effort measures (e.g. TOMM, MSVT, WCT, CVLT-FC, and RDS) independently. The statistics calculated included: hit rate, sensitivity, and specificity. The positive predictive powers (PPP) and negative predictive powers (NPP) of these measures could not be calculated despite recognizing the important contributions they would add to clinical decision-making. It is understood that these latter two indices incorporate the base rate of a condition of interest into classification models, and that failing to account for the influence of ecologically valid prior probabilities may result in unreasonably confident assertions regarding the accuracy of classification statistics (Rosenfeld, Sands, & Van Gorp, 2000), such as those

assessed by this aim. However, in order to adequately power all subsequent analyses (see Specific Aim 3) given the number of participants enrolled in the study, construction of criterion groups that reflected the 40% base rate of malingering had to be forfeited. Alternatively, the prior probability of suspect effort was artificially set at 50% to meet parameter demands (i.e., avoid overfitting the model): a decision that unavoidably nullified the utility of calculating PPP and NPP statistics in the present design.

The second objective of Aim 1 involved exploring each measure's unique ability to discriminate between the TBI and SIM criterion groups. Five logistic regression analyses were run, with each model employing one SVT as the covariate and group membership as the outcome variable. The resulting Nagelkerke R^2 (pseudo-R^2) was used as an approximate measure of the variance accounted for by each SVT in predicting group membership by assessing the associated correlations between the predictor(s) and criterion variables (i.e., the degree to which the model parameters improve upon the prediction of the null model), with high values being desirable. Receiver operating characteristics (ROC) were examined as another means of assessing the diagnostic efficiency of each SVT. ROC curve analysis generates information about discrimination capabilities of each model via the area under the curve (AUC), which ranges from .50 to 1.0 (high values indicate good discrimination). Model fit was also evaluated and compared using Bayesian Information Criterion (BIC) statistics, with more negative values being desirable.

Specific Aim 2: Examine the concordance among the SVTs. Specific Aim 2 examined concordance rates between classifications made by each SVT. The interrelationships among the measures were examined via correlational analysis of both the continuous and binary (pass/fail) outcomes of the measures, using tests appropriate to the level of the data (e.g., Pearson, point-biserial, etc.). Sets of concordance tables, trifurcated by group (e.g., TBI, SIM, and total sample),

organized the rates of agreement between all possible pairwise iterations of the SVTs.

Specific Aim 3: Determine combinations of SVTs that produce the most efficient, diagnostically valid index of suboptimal effort. Binary logistic regression served as the primary analytic strategy for Aim 3, which sought to discern combinations of SVTs that produced the most efficient, diagnostically valid index of suboptimal effort. All combinations of logistic regression models were tested, with the SVTs entered as covariates and group membership (TBI vs. SIM) as the outcome variable. These various models were then evaluated for their fit using multiple methods. AUC statistics for each model were compared. The Nagelkerke R^2 from the logistic regression models provides a useful index of variance explained by each model tested. Hosmer-Lemshow (H-L) statistics were employed to assess the calibration of the model, with non-significant values indicating adequate calibration. Model fit was also evaluated and compared using Bayesian Information Criterion (BIC) statistics, with more negative values being desirable.

Power analysis. The analyses were powered on the most demanding of the statistical tests, the logistic regression. A main concern in logistic regression is to avoid model overfitting. According to Harrell (2001), when a model is overfitted "it has too many parameters to estimate the amount of information in the data," and "the worth of the model will be exaggerated and future values will not agree with predicted values" (p. 60). A common cause of overfitting is employing too many covariates (predictor variables) relative to the number of cases. On the basis of models validated on independent datasets and simulation studies, sample size requirements are formulated as events per variable (EVP). Several studies have shown that the minimum EVP for obtaining reliable predictions is 10 (Harrell et al., 1984; Harrell, Lee, & Mark 1996; Harrell, Lee,

Matchar, & Reichert, 1985). Group sizes in the present study exceeded a conservative n-to-k ratio of 10 cases per variable in prediction models.

CHAPTER 3

RESULTS

Aim 1. Diagnostic Validity and Classification Accuracy for Single-Variable Models

Test Performance Based on Published Cut Scores. Specific Aim 1 sought to evaluate the diagnostic validity for each SVT individually. Initial classification accuracy statistics were calculated for classifications made by the TOMM, RDS, WCT, MSVT, and CVLT-FC individually predicting group status, using the dichotomous pass/fail classification based on cutoff scores indicated in the respective manuals. Phi coefficients reflecting the associations between group membership and pass/fail status on each of the five SVTs as based on the recommended cutoff scores were as follows: TOMM ($\varphi = .41$, $p < .001$), MSVT ($\varphi = .35$, $p < .001$), CVLT-FC ($\varphi = .35$, $p < .001$), WCT ($\varphi = .28$, $p = .002$), and RDS ($\varphi = .10$, $p = .150$). Thus, four of the five indexes were significantly associated with group membership, showing medium effect size, whereas RDS was unrelated to group membership. Resulting calculations of the sensitivity and specificity of the measures evaluated shows that both statistics vary modestly from one test to the next. As can be seen in Table 3a, sensitivity to suboptimal effort was largest (52%) when using the MSVT alone, followed by the TOMM (48%). The RDS yielded the smallest sensitivity (33%). Conversely, specificity was maximized using the CVLT-FC (93%), whereas the RDS yielded the smallest specificity (75%). As can be further extrapolated from Table 3a, the CVLT-FC generated the smallest proportion of false positive errors misclassifying TBI participants (i.e., 1 – Specificity) at 7% whereas the RDS yielded the largest proportion of false positive errors (25%). Concerning the mitigation of false negative rates (i.e., 1 – Sensitivity), the MSVT only missed 48.3% of the SIM group whereas the RDS missed 66.7%.

Test Performance Based on Logistic Regression. Table 3a also shows the classification statistics, ROC curve analyses, and BIC fit statistics for the single-variable models using the published cut scores (i.e., dichotomous pass/fail test results). Table 3b provides the chi-square statistics testing the significance (reliability) of the logistic regression models, as well as the odds ratios for each model. Although logistic regression typically uses continuous scores provided by a test, these analyses evaluate the test performance as defined by the respective manuals and as the tests are used clinically. Each of these single-variable models, except for that one using the RDS published cut-score as a predictor, was significant at $p < .05$; nonetheless, the quality of the models varied widely.

Traditional logistic regression models and ROC curve analyses also were used to assess classification accuracy for each of the individual SVTs using the *continuous scores* for each test. As a reminder, it should be noted that models assessing the accuracy of the MSVT employed the continuous MSVTavg variable for reasons discussed previously. For each analysis, the logistic regression used group membership as the outcome variable and the SVT (continuous raw scores) as the predictor. As seen in Tables 4a and 4b, each of the single-variable models using continuous raw scores as predictors was significant at $p < .05$; again, the quality of the models varied in classification accuracy and model fit (i.e., AUC and BIC statistics).

A strong indicator of a logistic regression model's ability to discriminate between groups (i.e., model fit) is the AUC produced by the model. This statistic, derived by calculating the area under the Receiver Operating Characteristic curve, provides information about how well the predicted probabilities created by the regression model match the observed probabilities over the entire range of values. In other words, it acts as a graphical representation of how well the model correctly classifies those cases with or without a condition of interest. Larger AUC values

represent better discrimination. AUC values at 0.50 offer no discrimination. AUC values between 0.70 and 0.79 are "acceptable," 0.80 to 0.89 are "excellent," and values greater than 0.90 are considered "outstanding" (Hosmer & Lemeshow, 2000). Despite the utility of AUC models in showing discrimination capability, this statistic can be relatively insensitive to changes in model fit when multiple covariates (i.e., predictors) are entered into the model, regardless of the apparent (i.e., via sensitivity, specificity, and associated classification accuracy statistics) predicative strength of any one of the added covariates. As a result, supplementing the AUC with other tests of model fit is beneficial.

Bayesian Information Criterion (BIC) were therefore calculated to estimate the probability that the predictor variables included in each model are contributing a significant effect to the model's ability to discriminate group membership. In analyzing the relationship between the estimated log likelihood function of the model and the number of explanatory predictors included, BIC imposes a penalty for increases in the number of predictors added to the model. Ultimately BIC reduces the inflated likelihood of the model, in the unexplained variance in the outcome variable (e.g., group membership), when the model is overfitted (i.e., the model is overloaded with non-essential explanatory variables, or predictors). BIC values are lowest when a model can explain the outcome values using the least number of parameters; thus, smaller BIC values imply good model fit and parsimony. However, BIC values cannot be interpreted in isolation; rather, they are interpreted via relative discrepancies across models (i.e., large differences in BIC values indicate strong preference in favor of the model with the smaller BIC value). An absolute difference of 0 – 2 is considered a "weak preference," 2 – 8 a "positive preference," 8 – 10 is "strong," and a difference greater than 10 is "very strong" (Raftery, 1996).

Tables 3a and 4a present classification and model fit statistics for each of the models

based on the published cut-scores and the continuous scores, respectively. These include Hit Rate, Sensitivity, Specificity, Nagelkerke R^2 from the logistic regression, and ROC area under the curve (AUC and AUC Confidence Interval), and Bayesian Information Criterion (BIC) statistics. Comparison of the two tables illustrates differences between the diagnostic validities of the tests as used in the clinical setting per the manual and the potential validities across all values of the scores. *As might be expected, in general, the validity statistics for dichotomous cutting scores are less precise than those observed for continuous scores on the tests.*

Overall, when dichotomous predictors were employed, the best single-variable model in terms of overall hit rate, AUC, and model fit was the TOMM; although, the MSVT and CVLT were nearly equivalent with only a weak preference indicated for the TOMM. Comparisons of BIC statistics in Table 3a show "very strong preference" for each of the TOMM, MSVT, and CVLT models over WCT and RDS. The WCT model shows "strong preference" over RDS. A similar pattern was produced when examining the continuous scores (Table 4a). Analyses produced a "very strong preference" for each of the TOMM, MSVTavg, and CVLT models over WCT and RDS, and "positive preference" for WCT over RDS; however, the discrepancies between the strongest models were larger: "strong preference" for the TOMM and "positive preference" for the MSVTavg over the CVLT.

Logistic regression indicated that the TOMM, using the published cut-score, was a significant predictor of group membership, $\chi^2 = 21.30$, $p < .001$, Exp(B) = 7.95, 95% CI [2.97, 21.31]. Nagelkerke's R^2 for the model was .22. However, area under the curve (AUC) for the TOMM was .69, 95% CI [.59, .79], which does not meet the .70 criterion to be considered "acceptable" (Metz, 1978). Of note, *none* of the models examining validity of the tests' published cutting-scores yielded AUC .70 or better. The worst-performing test was the RDS,

which was not a significant predictor of group membership, $\chi^2 = 1.10$, $p = .295$, Exp(B) = 1.54, 95% CI [0.685, 3.44] and showed the worst hit rate (54%), Nagelkerke's R^2 (.01), and AUC (.54, 95% CI [.44, .65]).

Interestingly, as seen in Table 4a, when the continuous raw scores were used as predictors, the sensitivity reflected in the logistic regression models was largest among the individual SVTs when group membership was predicted by the RDS (62%), but at the expense of specificity (51%): the lowest value for this particular classification statistic among the individual SVTs. Specificity was highest at 83% when group membership was predicted by the TOMM. Sensitivity was smallest when using the WCT (48%). As can be extrapolated from Table 4a, the proportion of misclassified cases in the TBI group (i.e., false positive for suboptimal effort = 1 - specificity) range from 17% (TOMM) to 49% (RDS). Misclassified cases in the SIM group (i.e., false negative for adequate effort = 1 - sensitivity) ranged from 38% (RDS) to 52% (WCT). As shown in Table 4a, when modeling the continuous raw scores, the TOMM was once again the strongest predictor of group membership as demonstrated in its ability to yield the largest hit rate (68%), variance accounted for (Nagelkerke's $R^2 = 0.30$), and AUC (.74). The RDS performed worst in these domains; although, as a continuous predictor, it was shown to be a significant predictor of group membership, $\chi^2 = 6.30$, $p < .05$, Exp(B) = 0.82, 95% CI [0.69, 0.97].

Aim 2: Compare the clinical utility of the SVT in relation to one another.

Simple concordance. Table 5 presents concordance tables for all possible pairs of SVTs. As shown, the percentages of overall agreement between the five measures are supplied for the TBI and SIM groups separately, as well as for the Total Sample. Also relevant are Tables 2a-c, which provide the Spearman Rho intercorrelations among the effort indices.

Considering the importance of accurately classifying bona fide TBI, Table 5 shows that the TOMM and CVLT demonstrated the highest agreement of TBI cases (46 cases; 87%) passing both measures. Only 1 case (2%) failed both tests, resulting in these tests showing the largest agreement of overall classification (89%) when organizing TBI group members. In those who passed the TOMM, 94% also passed the CVLT; of those who passed the CVLT, 94% passed the TOMM. Conversely, the MSVT and WCT yielded the highest proportion of TBI cases being misclassified as showing suboptimal effort as evidenced by 6 cases (11%) failing both tests. Of the 43 cases that passed the MSVT, 40 (93%) passed the WCT. Inversely, of those who passed the WCT, 91% also passed the MSVT. Overall agreement in organizing TBI group members was lowest between the MSVT and RDS (37 cases; 69%), followed closely by RDS and WCT (39 cases; 70%, respectively).

Also important is the classification agreement between measures when categorizing cases from the SIM group. Overall agreement was greatest between the TOMM and MSVT (51 cases; 88%). 26 cases (45%) from the SIM group were classified as showing suboptimal effort (i.e., failed both tests) when using the TOMM and MSVT; however, 25 cases (43%) passed both the TOMM and MSVT. Of those who passed the TOMM, 86% passed the MSVT. 89% passed the TOMM of those who passed the MSVT. Overall classification agreement of the SIM group was lowest when using the TOMM and RDS in combination (45 cases; 58%). The TOMM and RDS agreed in classifying 12 cases (20%) from the SIM group as showing suboptimal effort whereas 23 cases (38%) showed adequate effort by passing both measures. Out of those who passed the TOMM, 74% passed the RDS and only 58% passed the TOMM of those who passed the RDS.

Unlike a study such as this, in which group membership is known a priori, clinicians must rely on the results of the SVTs administered to best predict case classification. As such,

agreement of the measures using the total sample (i.e., both groups combined) is also shown in Table 5. As can be seen there, overall classification agreement was largest between the TOMM and MSVT. These tests agreed on the overall classification of 97 cases (87%), with 67 cases (60%) passing both tests and 30 cases (27%) failing both tests. Of those passing the TOMM, 87% passed the MSVT; of those passing the MSVT, 93% also passed the TOMM. The lowest rate of overall agreement was equally demonstrated by the TOMM and RDS as well as the MSVT and RDS, both of which classified 65% of the total sample similarly. Between the TOMM and RDS, 62 cases (53%) passed both tests and 14 cases (12%) failed both tests. Only 76% of those who passed the TOMM also passed the RDS and 75% of those who passed the RDS also passed the TOMM. Between the MSVT and RDS, 56 cases (51%) passed both tests, 17 (15%) failed both. 78% of those who passed the MSVT also passed the RDS and 71% of those who passed the RDS also passed the MSVT.

Descriptive intercorrelations among the SVTs (continuous scores) as well as correlations between the demographic variables are presented for the SIM Group, TBI Group, and Total Sample in Tables 2a through 2c, respectively. For the TBI group, correlations between SVTs, injury severity, and time since injury (months) are also presented (Table 2b). The tables present Spearman correlations because the distributions of most SVT variables were badly skewed. As can be seen in the tables, intercorrelations among the SVTs for the SIM group (Table 2a) were stronger than were those observed for the TBI group (Table 2b). Excluding TOMM Trial 1, which is not formally used as an index for classification, the median correlation among the SVT indices was .69 (range .28 to .91, mean rho = .66) for the SIM group. In contrast, the median correlation among the indices for the TBI group is .26 (range .11 to .83, mean rho = .34).

Aim 3: Determine combinations of SVTs that produce the most efficient, diagnostically valid index of suboptimal effort.

Initially it was planned to test all 26 possible combinations of 2-, 3-, 4-, and 5-multivariable models between the five SVTs to identify members of the TBI or SIM groups. However, given the demand of multivariable logistic regression analysis regarding limits on collinearity, this plan was not feasible. Specifically, collinearity diagnostics revealed that the TOMM and the MSVT could not be simultaneously included in a multivariable analysis. Therefore, the eight models requiring both the TOMM and the MSVT were excluded from this set of analyses.

Tables 3a (logistic regression models using the published cut-scores) and 4a (traditional logistic regressions using continuous scores) present the results of all models tested. Each table presents classification accuracy statistics (Hit Rate, Sensitivity, and Specificity), a measure of variance accounted for in terms of the association strength between the criterion variable and the various predictors (Nagelkerke's R^2), ROC curve analyses (AUC values and 95% confidence intervals), model calibration (e.g., Hosmer-Lemeshow χ^2, in which significant results indicate poor model calibration), and Bayesian Information Criterion (BIC) statistics. Tables 3b and 4b provide the chi-square statistics testing the significance (reliability) of the logistic regression models, as well as the odds ratios for categorical and continuous models, respectively.

Logistic Regression Derived Classification Accuracy Statistics – Two-Variable Models

As seen in Tables 3b and Table 4b, all two-variable logistic regression models tested were significant at $p < .05$; however, the models varied widely across the statistics that indicate relative quality. Initial comparisons began with observing differences in classification accuracy

between the models produced with categorical predictors and those created with continuous predictors. As shown in Table 3a, sensitivity calculated using published cut-scores ranged from 37% (RDS*CVLT model) to 55% when group membership was predicted with the MSVT*CVLT model. Specificity using these cut-scores ranged from 77% (MSVT*CVLT) to 93% using RDS and CVLT as predictors. Extrapolating from these statistics, it can be seen that the latter model yielded the lowest proportion of false positive errors (7%) and the MSVT*CVLT model generated the largest number of false positive errors (23%). Conversely, the RDS*CVLT model misclassified the largest proportion of SIM cases (i.e., False negatives = 63%) whereas the MSVT*CVLT produced the lowest proportion of these errors (45%).

As seen in Table 4a, the lowest sensitivity for continuous, two-variable models was derived by the RDS*MSVTavg model (49%). Both the RDS*WCT and WCT*CVLT models produced the largest sensitivity (55%). Compared against the dichotomous models, seven of the two-variable continuous models showed increases in sensitivity, two decreased, and the TOMM*CVLT remained the same (53%). Given the dynamic nature of the relationship between sensitivity and specificity, decreases in false negative rates (i.e., 1 – sensitivity) were coupled with moderate increases in the proportions of misclassified TBI cases (false positive rate; 1 – specificity): Specificity was lowered from a maximum of 93% ($RDS*CVLT_{categorical}$) to a high of 86% ($TOMM*RDS_{continuous}$ and $TOMM*WCT_{continuous}$). The lowest specificity obtained via the continuous, two-variable predictor models was 67% (RDS*WCT).

Logistic Regression Model Comparison – Two-Variable Models

All two-variable logistic regression models were significant. Among the categorical predictor models, only the MSVT*CVLT model found both tests to be significant predictors of group membership as seen in Table 3b. None of the continuous predictor models showed more

than one significant predictor (Table 4b).

Variance accounted for by the models varied modestly as shown in the ranges of Nagelkerke's R^2 presented in Table 3a and Table 4a. Regardless of the nature of the predictor variable (i.e., categorical vs. continuous), the models that included the TOMM and CVLT consistently showed a stronger association between the predictors and group membership. However, the continuous TOMM*CVLT model (Nagelkerke's R^2 = 0.33) outperformed the categorical TOMM*CVLT model (Nagelkerke's R^2 = 0.28). Among all two-variable models, this relationship was smallest when using the categorical RDS*WCT model (Nagelkerke's R^2 = 0.10); a result that remained consistent despite a minor increase when continuous RDS*WCT scores were employed (Nagelkerke's R^2 = 0.13).

Comparisons of the model fit of the categorical and continuous two-variable models, as seen in Table 3a and Table 4a, show that AUC was largest in the continuous TOMM*WCT model (AUC = 0.77, 95% CI [0.69 – 0.85]). Inversely, the model using the published cut-scores for the RDS and WCT yielded the weakest degree of discrimination of all two-variable models tested (AUC = 0.63, 95% CI [0.53 – 0.73]). A comparison of the patterns of discrimination performance across all models indicates that all but one continuous model (e.g., WCT*CVLT) yielded at least minimal to modest increases in AUC as opposed to those derived from the categorical models.

In the present analyses, the Hosmer-Lemeshow chi-square test for model calibration was used. A model is better calibrated when the observed and expected frequencies of group membership (as based on the predicted probabilities) are similar; therefore, non-significant (i.e., $p \geq .05$) differences are desired and indicate good fit of the model. As shown in Table 3a and 4a, the Hosmer-Lemeshow chi-square and *p* values for each of the regression models are provided.

All categorical and continuous two-variable models showed non-significant results, except for the continuous WCT*MSVTavg model (H-L $\chi 2 = 16.15$, $p = .04$).

Lastly, in order to assess the contributing effects of the predictors added to the models, BIC statistics were employed. Initial model comparison involved differentiating the BIC values between regression models using two-variable categorical predictors (Table 3a). The model most preferred was the TOMM*CVLT (BIC = -413.68). Degree of preference for this model is contingent on the BIC value of the model chosen for comparison. Thus, it can be said that there is a "very strong" preference (i.e., absolute difference = 22.82) in favor of the TOMM*CVLT model when compared to the model with the largest BIC, RDS*WCT (BIC = -390.86). However, there is only a "weak" preference (i.e., absolute difference = 1.91) when the TOMM*CVLT model is compared to the model with the second smallest BIC, the MSVT*CVLT (BIC = -411.77). In order to identify the strongest categorical model (overall), AUC values were used to first determine whether models were eligible for comparison; specifically, those with AUC values equal or greater than .70 (i.e., "acceptable") were compared. As shown in Table 3a, the categorical predictor model with the largest AUC was the TOMM*CVLT (AUC = 0.72), followed by TOMM*RDS (AUC = .70). Ancillary comparisons relying on BIC values indicate that there is a "very strong" preference (i.e., absolute difference = 11.46) in favor of the TOMM*CVLT (BIC = -413.68) over the TOMM*RDS model (BIC = -402.22).

Turning to the continuous predictor models, Table 4a shows that the range of AUC values is relatively constricted between the models that include the TOMM as a predictor. Amid these specific models the TOMM*WCT produced the largest AUC (0.77), with the next largest AUC being produced by the TOMM*CVLT (0.76). Despite the minor variations in the

discriminative ability of these models, BIC values indicates a "positive preference" in favor of the TOMM*CVLT model over the TOMM*WCT. Interestingly, despite the smaller AUC values of the MSVTavg*CVLT (0.73) and RDS*MSVTavg (0.73), BIC comparisons suggest that there is only a "weak" preference in favor of the TOMM*WCT over the MSVTavg*CVLT model, yet, a "strong preference" over the RDS*MSVTavg model.

Taking BIC into account while comparing all two-variable models, Table 3a and Table 4a highlight that the continuous TOMM*CVLT model is either positively or very strongly preferred to all other two-variable models, except the MSVTavg*$CVLT_{continuous}$. A point of particular interest was found when comparing of the same two SVTs using either the published cut-scores (e.g., TOMM*$CVLT_{categorical}$ BIC = -413.68) or the continuous raw scores (e.g., TOMM*$CVLT_{continuous}$ BIC = -418.44): results indicate a "positive" preference in favor of the continuous predictor model.

Logistic Regression Derived Classification Accuracy Statistics – Three-Variable Models

Three-variable combinations of the five SVTs were also entered simultaneously as predictors in logistic regression models. As seen in Table 3b and Table 4b, all models tested were significant at $p < .05$. Again, the resulting statistics varied considerably between models. As shown in Table 3a, sensitivity calculated using published cut-scores ranged from 48% (using the TOMM*RDS*WCT model) to 55% when group membership was predicted with either the RDS*MSVT*CVLT or the WCT*MSVT*CVLT models. Specificity using these cut-scores ranged from 76% (WCT*MSVT*CVLT) to 89% when using the TOMM, RDS, and WCT as predictors. Understandably, the TOMM*RDS*WCT yielded the lowest proportion of false positive errors (11%) and the WCT*MSVT*CVLT model generated the largest number of false positive errors (24%). The TOMM*RDS*WCT model misclassified the largest proportion of cases (i.e., False negative proportion = 52%) whereas the RDS*MSVT*CVLT and WCT*MSVT*CVLT models produced the lowest proportion of these errors (45%).

More often then not, sensitivity was improved using the continuous, three-variable predictors; although, in two instances (e.g., TOMM*RDS*CVLT and TOMM*WCT*CVLT) the models remained constant and the RDS*MSVTavg*CVLT model showed a reduction in sensitivity from 55% to 51%. Using the continuous scores for the TOMM, RDS, and WCT in combination increased the lowest categorically produced sensitivity from 48% to 52% and the largest categorical sensitivity increased from 55% to 56% (WCT*MSVTavg*CVLT). Again, recognizing the nature of the relationship between sensitivity and specificity, increases in sensitivity were coupled with decreases in specificity and subsequent increases in the rate of false positive errors. In particular, specificity was lowered from a maximum of 89% ($TOMM*RDS*WCT_{categorical}$) to 87% ($TOMM*WCT*CVLT_{continuous}$) resulting in a 2.2%

increase in false positive errors. The lowest specificity obtained via the continuous, three-variable predictor models was 79% (RDS*WCT*$CVLT_{continuous}$).

Logistic Regression Model Comparison – Three-Variable Models

As can be seen in Table 3b, among the categorical, three-variable models, only the RDS*MSVT*CVLT model resulted in more than one SVT being a significant predictor of group membership: the MSVT was a significant predictor with an odds ratio of 2.96 (95% CI = 1.11 – 7.94) as was the CVLT with an odds ratio of 4.02 (95% CI = 1.10 – 14.65). None of the models using continuous predictors showed more than one significant independent variable (Table 3b).

The variability of the values for Nagelkerke's R^2 were modest, as shown in the ranges presented in Table 3a and Table 4a. Identical to the two-variable models, any three-variable model that incorporated the TOMM and/or CVLT as predictors consistently produced the largest predictor to criterion associations, regardless of the nature of the predictor variable (i.e., categorical versus continuous). Moreover, the models using the continuous scores from both of these SVTs as predictors resulted in larger Nagelkerke's R^2 than any of the categorical TOMM*CVLT models. For example, between all three-variable models tested, the Nagelkerke's R^2 was largest for the TOMM*WCT*CVLT (0.35) model using continuous predictors whereas the same model using categorical predictors resulted in a Nagelkerke's R^2 of 0.28. Across all three-variable models, Nagelkerke's R^2 was smallest when using the categorical RDS*WCT*MSVT model (0.16).

AUC values for all three-variable models were compared as in the two-variable models. Again, these AUC values were supplemented with the Hosmer-Lemeshow chi-square test for model calibration (see Tables 3a and 4a). All categorical and continuous two-variable models

showed non-significant results; therefore, all models were adequately calibrated.

Comparisons of the values for AUC listed in Table 3a and Table 4a again show the largest (although, again only "acceptable") AUC being produced by the continuous TOMM*WCT*CVLT model (AUC = 0.78, 95% CI [0.69 – 0.85]); a minor increase as compared to the two-variable, continuous TOMM*WCT model (AUC = 0.77, 95% CI [0.69 – 0.85]). The RDS*WCT*MSVT, RDS*WCT*CVLT, and RDS*MSVT*CVLT produced equal discriminability (AUC = 0.68, 95% CI [0.59 – 0.78]) amid the three-variable, categorical models. However, AUC was lowest when the model used the continuous RDS, WCT, and CVLT scores as predictors (AUC = 0.67, 95% CI [0.57 – 0.77]). As noted in the two-variable model results, AUC values can be relatively insensitive to changes in model fit despite adding predictively strong independent variables to the model. Comparing the two-variable AUCs to the three-variable AUCs (respective of predictor type) highlighted this point by showing minimal increases in discrimination despite the addition of a third predictor to each model.

Direct model comparisons involved identifying the smallest BIC values (Table 3a) within regression models using two-variable categorical predictors, then within the two-variable continuous models, and lastly between all three-variable models. As before, eligibility for BIC comparisons required at least an "acceptable" AUC value (e.g., AUC >= 0.70). Among the categorical models, the TOMM*WCT*CVLT was most preferred as measured by its production of the smallest BIC (-409.86) and largest AUC value (0.73). Following, the TOMM*RDS*CVLT model yielded an AUC of 0.71 and a BIC value of -408.92. The differences in the BIC values between these models only produced a "weak" preference in favor of the TOMM*WCT*CVLT.

Similar to the two-variable continuous models, Table 4a shows that the ranges of AUC

values produced by the three-variable models using continuous predictors was relatively small. Despite minimal difference between the largest AUC value (TOMM*WCT*CVLT = 0.78) and second largest AUC value (TOMM*RDS*WCT = 0.77), the difference between the BIC values (-415.85 and 408.18, respectively) points to a "positive preference" in favor of the TOMM*WCT*CVLT. Considering all BIC values listed in Table 3a and Table 4a, the latter model is the most preferred model among all three-variable combinations and is considered to have a "positive" preference as compared to the same model using categorical predictors.

Logistic Regression Derived Classification Accuracy Statistics – Four-Variable Models

Next, the four-variable combinations of the five SVTs were used as predictors in logistic regression models. Table 3b and Table 4b shows all four models were significant at $p < .05$. Classification accuracy statistics calculated using published cut-scores are listed in Table 3a. The TOMM*RDS*WCT*$CVLT_{categorical}$ model produced 53% sensitivity and 87% specificity. The RDS*WCT*MSVT*$CVLT_{categorical}$ model obtained a sensitivity of 55% and specificity of 76%. The former model produced a slightly greater proportion of false negative errors (47%) as compared to the latter model (45%). However, the inverse was true of the potentially more detrimental false positive errors whereby the TOMM*RDS*WCT*$CVLT_{categorical}$ produced a smaller proportion (13%) of these errors than did the RDS*WCT*MSVT*$CVLT_{categorical}$ (24%).

Improvements in sensitivity were produced when the test's continuous scores were used as predictors in the four-variable models. The sensitivity of the TOMM*RDS*WCT*$CVLT_{continuous}$ model increased to 55% and the RDS*WCT*MSVT*$CVLT_{continuous}$ increased its sensitivity to 56%. Conversely, specificity decreased to 85% when the continuous predictors were used in the TOMM*RDS*WCT*CVLT

model; resulting in a slight increase in the proportion of false positive errors (15%) when using this model. Interestingly, specificity increased to 78% when the RDS*WCT*MSVT*CVLT used continuous predictors.

Logistic Regression Model Comparison – Four-Variable Models

Tables 3b and 4b show that none of the models using either categorical or continuous predictors produced more than one significant independent variable. Once more the same pattern of significant predictors reported in all preceding models held constant for all the four-variable models; the TOMM was the only significant predictor of group membership in the TOMM*RDS*WCT*CVLT models and the CVLT was solely significant in the RDS*WCT*MSVT*CVLT models. Of special note, the RDS proved to be a non-significant predictor in all four-variable models that incorporated this SVT.

The pattern of Nagelkerke's R^2 values produced in the four-variable models held constant with the patterns observed in all preceding models. As can be seen in Table 3a and 4a, the models using the continuous scores as predictors resulted in larger Nagelkerke's R^2 than any of the categorical models. Specifically, the four-variable model yielding the largest Nagelkerke's R^2 was the TOMM*RDS*WCT*CVLT (0.35) when continuous predictors were included as compared to 0.28 when categorical predictors were used. The categorical RDS*WCT*MSVT*CVLT model produced the smallest Nagelkerke's R^2 (0.22) among all the four-variable regression models.

AUC values for all four-variable models can be found in Table 3a and 4a. The Hosmer-Lemeshow chi-square and *p* values for each of the four-variable regression models show that all four-variable models were at least adequately calibrated as evidenced by the absence of *p* value

less than or equal to 0.05. 75% of the four-variable models produced "acceptable" AUC values, with the largest being generated by the continuous predictor, TOMM*RDS*WCT*CVLT model (AUC = 0.79, 95% CI [0.71 – 0.81]). The smallest AUC value was produced by the RDS*WCT*MSVT*$CVLT_{categorical}$ model (AUC = 0.69, 95% CI [0.59 – 0.79]).

Of the two categorical models, BIC was smallest (-405.09) for the TOMM*RDS*WCT*CVLT model as compared to the RDS*WCT*MSVT*CVLT model (BIC = -403.55). The direct comparison of these two models resulted in a "weak" preference in favor of the TOMM*RDS*WCT*CVLT. Similarly, the continuous TOMM*RDS*WCT*CVLT model produced the smallest BIC value (-411.59) as compared to that produced by the continuous RDS*WCT*MSVTavg*CVLT (-410.65). The magnitude of the difference between these BIC values points to a "weak" preference in favor of the TOMM*RDS*WCT*CVLT.

Model Comparisons between All Regression Models

The five-variable model was deemed unstable due to the necessary inclusion of both the TOMM and MSVT, which when used in conjunction produced unacceptable levels of collinearity. Thus, this model was excluded from the present analyses. Therefore, the final analyses inspected the direct comparisons of all eligible models against each other, regardless of the number of predictors included.

First, models using only categorical predictors were evaluated by comparing the BIC values between models producing the largest "acceptable" AUC values. Within these models, AUC values were equally large (0.73) for the three-variable TOMM*WCT*$CVLT_{categorical}$ and the four-variable TOMM*RDS*WCT*$CVLT_{categorical}$ models. The former model yielded a BIC of -409.86 whereas the latter generated a BIC of -405.09. The magnitude of the difference

between these scores was 4.77, indicating a "positive" preference in favor of the three-variable model. Although the two-variable TOMM*CVLT model yielded a slightly smaller AUC (0.72) than the previous models, its BIC value of -413.68 suggests a "positive" preference in favor of this model over both previously noted categorical models.

Between the models using continuous predictors (Table 4a), the TOMM*RDS*WCT*CVLT model achieved an AUC value of 0.79, the largest of any continuous predictor model. This model was followed by the TOMM*WCT*CVLT (AUC = 0.78), TOMM*RDS*WCT (0.77), TOMM*CVLT (0.76) & TOMM*RDS*CVLT (0.76), TOMM*RDS (0.75), TOMM (0.74), and RDS*MSVTavg, MSVTavg*CVLT, RDS*MSVTavg* CVLT, RDS*WCT*MSVTavg*CVLT (each of which obtained an AUC of 0.73). Of these, the two-variable TOMM*CVLT model produced the smallest BIC value (-418.44), followed by the MSVTavg*CVLT model (BIC = -417.25), the TOMM*WCT*CVLT (BIC = -415.85), the single-variable TOMM model (BIC = -415.00), and the TOMM*WCT (BIC = -412.55). Based on BIC statistics alone, the $TOMM*CVLT_{continuous}$ model was preferred over all the continuous predictor models; however, the degree of preference was still considered "weak" as compared to the MSVTavg*CVLT model and only a "positive" preference over the TOMM in isolation.

Furthermore, the TOMM*CVLT yielded the smallest BIC among all models tested (i.e., any model using either categorical or continuous predictors). For example, the largest categorically derived AUC value (0.73) was obtained by both the TOMM*WCT*CVLT and TOMM*RDS*WCT*CVLT, however, their respective BIC values (e.g., -409.86 and -405.09) indicate a strong or very strong preference for the continuous TOMM*CVLT model. Even more, the continuous TOMM*CVLT model proved to be "positively" preferred over the categorical TOMM*CVLT model (AUC = 0.72, BIC = -413.68).

Although some of the categorical models produced relatively larger rates of specificity (and subsequently fewer false positive errors) than the continuous models were able to provide, the majority of these categorical models had unacceptable values of AUC and BIC statistics as compared to the continuous TOMM*CVLT model. For example, although the categorical RDS*CVLT yielded a specificity of 93% as compared to the continuous TOMM*CVLT model's 85% specificity, the former only produced an AUC of 0.67 and BIC of -401.36. Despite producing a smaller proportion of false positive errors (7%), the $RDS*CVLT_{categorical}$ model generated a less than "acceptable" AUC value and a BIC that indicated a "very strong" preference for the $TOMM*CVLT_{continuous}$ model. Of those categorical models that produced "acceptable" AUC values and larger specificities than the continuous TOMM*CVLT model (e.g., TOMM*RDS, TOMM*RDS*CVLT, and TOMM*RDS*WCT*CVLT), none generated BIC values small enough to indicate a preference over the continuous $TOMM*CVLT_{continuous}$.

CHAPTER 4

DISCUSSION

The findings partly support the primary hypothesis that using multiple measures of effort increases classification accuracy when discriminating between bona fide and feigned traumatic brain injury (TBI); however, not nearly as strongly nor as ubiquitously as had been presumed. Contrary to initial expectations based on the patterns of performance theory, these findings support the position that the incremental validity generated by continually adding effort measures to a battery is not unlimited. Rather, given the five symptom validity tests (SVT) compared, the accuracy of correctly identifying group members showed only minor to modest growth when the number of tests modeled was increased. To a large degree, this finding reflects the strength of the Test of Memory Malingering (TOMM), Medical Symptom Validity Test (MSVT), and California Verbal Learning Test – Forced Choice Trail (CVLT) as individual predictors among this specific set of tests, inasmuch as adding meaningfully to a pair of these tests proved difficult. Otherwise stated, these three SVTs ubiquitously outperformed others and they were relatively equivalent in terms of their ability to distinguish between group members and they were largely concordant in their decisions. Therefore, adding any of these three measures to another symptom validity test (e.g., the Word Choice Test or Reliable Digit Span) generally improved the predictive accuracy and fit of the decisional model, regardless of the model's clinical or statistical orientation.

The findings also highlight an important psychometric distinction between examining classification accuracy of the SVTs as used clinically (i.e., categorical cutoff scores identified in the test manuals) and as traditionally tested statistically (i.e., continuous raw scores). As expected by theory, predictive powers of continuous scores on the SVTs were greater than for

categorical scores on the SVTs; however, models tested using continuous scores overestimate the ecological validity of the tests as used in the clinical setting with pass/fail scores.

There is no single, "gold standard" method of evaluating the multifaceted dimensions of a test's (or battery's) decisional accuracy. Rather, the dynamic and paradoxical relationship between the indices of basic classification accuracy (e.g., hit rate, sensitivity, and specificity) and the statistical methods for assessing models produce a cornucopia of information that do not inherently coalesce into an absolute standard for evaluating a measure or battery. Instead, this plethora of information must be teased apart, weighted, and clinically implemented on a case-by-case basis. Nevertheless, these findings produced a number of invaluable pieces of information that support the efficient and clinically useful process of assessing measures of symptom validity in populations affected by TBI.

Considering the clinical and pragmatic importance of deriving a parsimonious battery, the findings indicate that using the TOMM and CVLT in conjunction maximized predicative accuracy as compared to any other single or assortment of effort measures. The combined effects of the MSVT and the CVLT produced a very similar outcome. Opposite to initial presuppositions, the findings demonstrated that little is gained in terms of decisional accuracy when including more than these pairs of sound SVTs; a conclusion similar to that drawn by Victor et al. (2009) who stated that a two-measure, pairwise failure model acts as the optimal criterion (i.e., in terms of clinical utility and parsimony) for identifying non-credible performance. Moreover, these findings highlight that the inclusion of suboptimal tests can increase the potential for unnecessary examinee fatigue, or worse, detract from the validity of a battery's classification accuracy. In particular, these findings show that not only did the Reliable Digit Span (RDS) fail to improve any model to which it was added, but also in many cases it

appeared to adversely affect the decisional models when employed.

Overall, the lack of a large, reliable increase in decisional accuracy when combining three or more of these widely used SVTs highlights the distinction between statistical (i.e., theoretical) psychometrics versus clinically applied psychometrics. The quantitative improvements generated by administering additional measures may be more a sign of statistical illusion than clinical significance. Taken together, these findings support the conclusion that predicative accuracy is technically and theoretically strengthened when multiple measures of effort are given, yet, the clinical administration of any more than two of the most psychometrically sound measures results in either redundancy or, in some cases, diminished clinical utility and predicative accuracy.

Specific Aim 1: Diagnostic validity of the individual SVTs

The hypothesis that each of the effort measures used in this study would demonstrate sound psychometric properties was generally supported. Of note, various psychometric indices of goodness did not always agree; for example, on some occasions, classification accuracy statistics favored one SVT, whereas indices of parsimony favored another. Overall, considering the five SVTs individually, the TOMM produced the strongest profile of diagnostic validity, with multiple indices favoring the TOMM over the other SVTs. The MSVT was second strongest of the SVT set. Overall, both the MSVT and CVLT obtained a hit rate of 66%, whereas the TOMM correctly classified 68%. The RDS and the WCT performed relatively poorly as compared to these three tests; although sensitivity of the continuous model RDS was highest.

Amid all SVTs tested, classifications made using a dichotomous pass/fail score generated fewer false positive errors than classifications based on continuous scores: the continuous models consistently yielded poorer specificity. This finding is in line with the assumption that the test

developers' would publish cutoff scores that maximize specificity due to false positive errors being considered more egregious misclassifications in most clinical settings.

The RDS showed the largest decrease in specificity when its continuous data were used; specifically, moving from a respectable 75% using the published cutoff score to an unacceptably low 51%. Given the inverse relationship between specificity and sensitivity, the decreases in the specificities of continuous models were accompanied by larger sensitivities than those generated by the categorical models. Altogether, the findings suggest that each SVT is capable of generating adequate, yet conservative, classifications when used as instructed by the publishers. Conversely, models evaluated using the continuous scores yielded larger, unacceptably high, probability of false positive errors; however, this risk was typically rewarded with a greater percentage of overall correct decisions (i.e., hit rate). It seems important to note that evaluation of the tests using continuous scores represents a theoretical entity (i.e., examining the full range of possible cutting scores) and not an applied rule, as is employed in clinical settings.

Specific Aim 2: Comparative clinical utility of the SVTs.

Heuristically, relying on one piece of data when making a clinical judgment is not acceptable under any circumstance. This rule-of-thumb is especially important in the context of medico-legal forensic evaluations where diagnostic errors may result in a patient's loss of freedom, access to care, or the unfair disbursement of monetary compensation. As suggested by Millis (2010), a number of steps can be implemented to decrease the potential for erroneous decision making: incorporating the use of base rates, seeking out disconfirming evidence, and allowing statistics to inform decisional rules. Given the wide assortment of SVTs available to a clinician assessing for poor effort, this study aimed to incorporate the latter suggestion of using statistics as a way to empirically guide the construction of an incrementally valid effort battery.

Relying on the framework of the patterns of performance theory, it was hypothesized that a battery of multidimensional, multi-method SVTs showing low to moderate intercorrelations would not only produce increases in the detection rate of poor effort as compared to isolated measures, but also identify the most incrementally valid set among the 26 possible combinations of the five SVTs.

Incremental validity, as defined by Haynes and Lench (2003), is "the degree to which an instrument provides measures that are more valid than alternative measures of the same variable." Mathematically, incremental validity is the proportional increase of correct decisions that result from using one test or battery over another (e.g., positive predictive power of the battery minus the base rate of the condition of interest). The base rate of malingering in TBI-related cases varies by clinical setting; however, it is generally accepted that 30 – 40% of mild TBI cases entering a medico-forensic setting involve malingered neurocognitive deficits. Unfortunately, calculating the predictive powers of the models using a base rate of 40% was not possible due to restrictions in sample size and the demand characteristics of using logistic regression models (e.g., events per variable restrictions required the use of all subjects to avoid overfitting models that incorporated each of the SVTs simultaneously). Therefore, the base rate was artificially set at 50% to avoid "overfitting" the models when all five SVTs were included.

Given the inability to calculate incremental validity in the traditional sense, other criteria for comparing the SVTs were employed. These included evaluating changes in hit rate, sensitivity, and specificity as well as using statistical methods to compare the calibration, fit, and discrimination capabilities of the logistic regression models. However, determining model preference based on these characteristics comes with the inherent problem of assigning subjective weights to each statistic. There is no standard method for rank ordering the

importance of hit rate versus specificity, for example, or AUC versus BIC values. Ultimately these decisions boil down to the opinion of the clinician who must appraise the importance of the model statistics as they apply to his or her population of interest and clinical needs. For the purpose of this study, AUC and BIC values were ranked as most important, followed by hit rate, specificity, and sensitivity: a decision that was based on statistical pragmatics.

The importance of AUC was prioritized as a method for determining the optimal "balance" between maximizing sensitivity without diminishing the model's ability to reduce false positive errors (i.e., 1 – specificity). Given that the dynamic relationship between sensitivity and specificity (both of which affect hit rate) is contingent upon both the test or battery administered and clinical judgment concerning their relative importance or appropriateness to a clinical setting, decisions about model superiority relied heavily on receiver operating characteristics (ROC) and area under the curve (AUC) generating "acceptable" (AUC $\geq$ 0.70) equilibrium between these two classification accuracy statistics. Although this statistic provided a useful measure of group membership discrimination between the models, its insensitivity to the addition of predictor variables required a supplemental criterion. Hence, Bayesian Information Criterion (BIC) statistics were incorporated to deduce the significance of each predictor's contribution to the model. Importantly, however, was the decision to grant hierarchical primacy to AUC over BIC, as BIC is prone to "rewarding" model parsimony. Although keeping assessment batteries short is clinically pragmatic, this study was focused on the incremental validity of multi-method patterns of performance. Hence, small BIC values generated as a result of the brevity of a battery were not *necessarily* given qualitative superiority over longer batteries that produced large AUC's, hit rate, sensitivity, or specificity.

Unfortunately, no gold standard of symptom validity or effort exists. As a result,

determining the conclusive validity of SVTs is not possible; rather, the clinician must rely on the simple concordance or reliability between measures (Axelrod & Schutte, 2011). The SVTs showed medium to strong associations with each other. Notably, however, if multiple measures are used, clinical utility is only increased if there is minimal shared method variance between the tests (Meyer, 2003). However, modest associations should be expected between tests that purport to measure the same construct (e.g., effort), in the same cognitive domain (e.g., memory), using a similar methodological paradigm (e.g., forced-choice). Given these similarities, evidence of convergent validity is desirable and was observed in the modest interrelationships found among the SVTs. This finding suggests that these tests are likely measuring the same construct. Still, it is important to consider those factors (other than shared methods variance) that may be influencing the associations between the SVTs. Specifically, attenuation of the associations between SVTs may occur because of sensitivity to constructs other than effort (i.e., they are not as robust as assumed) or they are picking up on different facets of a multidimensional effort construct.

Concerning the first point, regarding an SVT's sensitivity to constructs other than effort, the findings indicate that at least one of the SVTs employed (i.e., the RDS) like demonstrates over-sensitivity. For example, approximately 25% of bona fide TBI cases failed the RDS; however, 75% of those who failed passed every other SVT. When examining TBI and SIM participants combined, roughly 30% failed the RDS: 38% of these were TBI cases. Of those who failed the RDS, 41% passed all four other measures and 75% passed at least one other SVT. The TBI cases accounted for 69% of those passing all four SVTs as well as 50% to 63% of those passing between one and three other test, respectively. Strikingly, after having failed the RDS, not a single TBI participant failed all four of the remaining SVTs, whereas 40% of simulators

did. Concordance rates such as these strongly demonstrate that a TBI survivor's failure on the RDS does not well predict performance on the other SVTs. Moreover, compared to the range of TBI failure rates seen on the other four effort measures (e.g., 7.5 – 18.5%), the 25% failure rate on the RDS represents a 35 - 233% increase in bona fide TBI survivors' being classified as showing insufficient effort based on this test. These discrepancies suggest that the RDS is likely picking up on a trait dissimilar to that being measured by the other effort measures. The second point, that the construct of effort may be multidimensional, has very important implications for the use and interpretation of SVTs in general. Most notably, false positive rates can be unduly inflated by the presence of either related (e.g., intention vs. effort) or non-related (e.g., symptoms of affective disorders) constructs (Frederick, 2009). Of all the measures tested, the RDS consistently yielded the smallest intercorrelations with the other tests, ranging from .28 with the MSVT to .47 with the CVLT. The RDS also generated the smallest hit rate and specificity regardless of variable type, and conversely, it produced the largest sensitivity when its continuous scores were modeled. Consequently, the RDS yielded the largest number of false positive errors of any SVT assessed. Considering the structure of the test, the RDS is the only measure of the five that does not utilize a forced-choice paradigm. Rather, obtained scores are derived strictly from non-cued performance. Thus, it may be that RDS's minimal relationship with the other measures is due to the lack of shared methods variance. However, it is maintained that the more likely possibility is that the RDS is overly sensitive to something other than "effort" as operationalized by the other SVTs (i.e., brain injury). Taken together, these finding may provide some evidence for an inherent problem with "embedding" a symptom validity test into a measure initially created to assess a more specific domain of cognitive ability (e.g., executive functioning or working memory); especially one to which TBI survivors are acutely

susceptible given the typical sequelae of moderate to severe injuries (Riggo, 2011).

An SVT's robustness to constructs other than effort is an extremely important characteristic of this type of measure, as it allows the administering clinician to infer that a positive test result is due to the absence of effort and not the presence of cognitive, emotional, or psychiatric problems. Generally, cognitive deficits following an uncomplicated mild TBI are not chronic; they typically resolve around one month post-injury. However, the neurobehavioral sequelae of complicated, moderate, and severe TBI can be varied and create lasting impairments across domains such as attention, memory, executive functioning, aggression, poor impulse control, anhedonia, or apathy (Riggo, 2011). Furthermore, depression is shown to occur in 25 – 50% of individuals after a moderate to severe TBI and is often accompanied by symptoms such as fatigue, distractibility, irritability, and rumination (Seel & Kreutzer, 2003). Although some SVTs have been repeatedly shown to be insensitive to these cognitive impairments (e.g., the TOMM), others lack this verification. As a result, failures across multiple SVTs are not uncommon for patients demonstrating verified cognitive symptoms (Merten et al., 2007). In the same study, Merten and his colleagues showed that performance on the RDS, in particular, is heavily influenced by cognitive impairment such as those seen in some moderate to severe TBI survivors. Although it is impossible to completely rule out poor effort in patients suffering from cognitive impairments, it seems much more plausible that their effort, as indicated by a SVT such as the RDS, is being categorized as suboptimal because of inadequate specificity. Given that all TBI participants in this study were verified as moderate or severe, it seems most likely that the globally poor performance of the RDS was heavily influenced by constructs other than effort.

On another hand, it may be that the RDS performed poorly due to the cutoff score used. Although a reliable digit span forward plus backward equaling 7 or less is the published cut point, recent research suggests this may be too stringent a cut score. Using ≤ 6 increases specificity while suppressing sensitivity, a result that would reduce the number of false positive errors made by the RDS (Babikian, 2006).

Specific Aim 3: Best combinations of SVTs.

Here the focus was on deriving a combination of SVTs that produced the most efficient, diagnostically valid index of suboptimal effort. It was hypothesized that incremental validity would be enhanced with the inclusion of measures using multiple, distinct methods for assessing effort. The findings of this study supported this hypothesis in part; however, the results obtained did not align with the patterns of performance theory in the manner that was expected. Specifically, the findings suggested that some SVTs were inherently stronger (psychometrically) and thus played a larger role in the accuracy of classification, other measures detracted from the model's classification accuracy, and most importantly, optimal models tended to include fewer measures as opposed to more. Overall, the assumption that more is better was overruled by findings suggesting that given the right measures, less is more.

A particularly striking pattern was observed for the combined predictive power of combinations of two SVTs. Any combinations that included the TOMM found it to be the only meaningful predictor of group membership. The same pattern held for any TOMM-absent pair that included either the MSVT or the CVLT; both the MSVT and CVLT outperformed the WCT and RDS in any pair in which either was included, and neither the WCT nor the RDS added meaningful predictive value when combined with TOMM, MSVT or CVLT. In fact, the RDS

was never a substantial predictor of any model, and the WCT outperformed the RDS whenever these tests were combined. These findings support the conclusion that the TOMM, MSVT, and CVLT are the strongest predictors of group membership within their respective batteries, and it also demonstrated that these three tests were remarkably equivalent in their classifications, as evidenced by the strong concordance between the measures. Overall agreement in classifying TBI cases was 89% for the TOMM and CVLT, 85% for the TOMM and MSVT, and 80% for the MSVT and CVLT. Overall agreement in classifying Simulator cases was 88% for the TOMM and MSVT, 79% for the MSVT and CLVT, and 78% for the TOMM and CVLT.

The pattern noted for pairs of SVTs was also evident in all combinations of three and four tests. However, despite the MSVT outperforming the WCT and RDS, the CVLT outperformed the WCT, RDS, and MSVT in the four-test combination including RDS, WCT, MSVT, and CVLT. Yet, surprisingly, both the MSVT and CVLT added meaningful predictive value in the categorical model including the RDS, MSVT, and CVLT, whereas the CVLT was the only meaningful predictor in the continuous RDS*MSVT*CVLT combination.

The findings also provided unique exploratory information concerning the effects of modeling categorical SVT classifications as opposed to those derived from continuous SVT scores. This theoretical issue would appear to have meaningful clinical implications, because in the clinical setting, clinicians apply a single cutting score as recommended by the test manual and do not benefit from the theoretical range of cutpoints as is tested in traditional statistical analyses. The findings showed that the use of continuous SVT scores frequently generate stronger results in terms of distinguishing group membership (e.g., larger AUC values) than those created using categorical variables. For example, given that an “acceptable” AUC of 0.70 or greater is desired, 65% of the models based on continuous scores met this criterion whereas

only 26% of the categorical models achieved this standard. Next, relying on BIC values as an index of the superiority of models, direct comparisons between categorical and continuous models consisting of the same SVTs indicate preference in favor of all continuous models. Taken together, these findings indicate that examining models using continuous scores may overestimate the actual performance of the tests as used in the clinical setting. It is not feasible to employ multiple cutpoints for a single case; thus, relying on the psychometric properties of models derived from continuous scores is inapplicable within the clinical context and changes in discriminability merely reflect conceptual (as opposed to pragmatic) comparisons. However, this understanding should not undercut the importance of relying on statistical methods (e.g., ROC curves) when initially identifying an ideal cutscore to use with a specific clinical population or setting.

Altogether, the best combination of SVTs is the TOMM in conjunction with the CVLT. With a combined hit rate of 0.68, a modest sensitivity of 0.53, and acceptable specificity of 0.85, this two-test battery offers a well-rounded accuracy in the most parsimonious package. Adding tests to this battery, such as the WCT or RDS, provide little improvement across these domains, and thus, the cost-reward ratio favors reducing administration time, examinee fatigue, and the probability of obtaining a false-positive error due to chance. Furthermore, the TOMM appears to have strong predictive ability in terms of performance on alternative measures of effort. In particular, the concordance of the TOMM with the MSVT, RDS, and CVLT shows that of those who fail the TOMM, 86% failed the MSVT, 61% failed the CVLT, and 40% failed the RDS. Rates such as these demonstrate that the addition of alternative measures to a battery including the TOMM would be likely redundant, offering only meager improvements in specificity while wasting alternative measures should retest be required. Concerning the CVLT, the results show

that it obtained the highest specificity, misclassifying 7% of the TBI group. The ability to be 93% certain when ruling out feigning is a particularly desired trait in an effort test. Hence, it is not surprising that the addition of this embedded measure to a battery including the TOMM would provide excellent classification accuracy. Overall, it appears that between the TOMM's adept hit rate (resulting from a well balanced ratio of sensitivity to specificity) and the CVLT's powerful avoidance of false positives, combining these two SVTs offer the most statistically sound model for predicting group membership.

Given that clinicians rely on the publication manuals for scoring and interpretation, the results obtained from generating theoretical models that examine the range of continuous scores (i.e., all possible cutting scores) are impractical. Also, although combinations of tests based on continuous scores yield models that are *statistically significant*, the resulting increases in accuracy may not be *clinically meaningful*. For example, despite the statistical advantages of adding the CVLT to the TOMM, the 1-point improvement to hit rate may be clinically negligible. Although statistically significant psychometric differences appear important on paper, they may bear little impact on clinical reliability or significance.

Given the near equivalent strength of the TOMM, MSVT, and CVLT, a striking clinical inference can be drawn. Ignoring both the patterns of performance theory and the adage that a clinical decision should not be made on one piece of data alone, it seems that remarkably little incremental validity is added beyond administering the TOMM in isolation. Furthermore, the MSVT and CVLT perform nearly as well when using the published cutoff scores. Thus, if parsimony or battery brevity is highly important, then these findings provide strong evidence for administering the TOMM alone, while reserving the MSVT or CVLT as equivalent, alternative measures for future assessment. For example, the findings show that the hit rate for the TOMM

only increases by 3 to 4 points when three additional measures are included in the battery. For that matter, sensitivity also appears remarkably unchanged despite the addition of multiple measures to the TOMM. And, as would be expected, it is specificity that jumps most reliably as the number of SVTs administered increase, yet even this improvement is modest at best. So, in stark contrast to the expectations born of the patterns of performance theory, for this specific set of SVTs, predicative accuracy is generally equivalent when selecting one of the most reliable, robust, and clinically efficient tools such as the TOMM, MSVT, or CVLT.

Provided that the clinician does not permit potential sacrifices to validity in favor of parsimony, however, these findings can also provide an empirically guided strategy to bolstering the validity of an effort battery. If it is decided that another test needs to be added to the battery, then it appears that supplementing any battery with the TOMM will yield improvements in decisional accuracy. Similarly, adding the MSVT or CVLT to any battery will improve the model fit and generally raises its classification accuracy. However, combining the TOMM and MSVT cannot be recommended, because these were so strongly related that combined models could not be reliably tested using multivariable statistics; in that regard, the strong concordance might indicate that little to no incremental validity will be obtained. The WCT or CVLT, on the other hand, given their inherently large specificity, will typically produce a clinically meaningful reduction in false positive errors if added to a battery. For example, a clinician is faced with an 11% chance of making a false positive error if the TOMM is failed; yet, requiring that the CVLT also be failed reduces this error rate to 1.9%. Ultimately, the only measure that detracted from the accuracy of clinical decision-making was the RDS. As noted, this problem may reflect its sensitivity to cognitive functioning above and beyond effort, and thus, the RDS may not be an appropriate measure of effort when assessing a survivor of TBI. In sum, of these five well-

known measures assessed, clinicians employing the TOMM and CVLT in combination will likely obtain the most valid results concerning their examinee's test-taking effort.

Limitations:

The most evident shortcoming concerned the sample recruited for the study. Although inclusion criteria required that the TBI sample consisted of well-verified moderate to severe cases in order to maximize experimental control, this degree of management resulted in costs to generalizability. In particular, the inclusion criteria were such that uncomplicated mild or very severe TBI cases did not participate; thus, these findings may not generalize well to discrimination of effort versus bona fide TBI for these subgroups. The extant literature reports that the base rate for feigned neurologic impairment is largest among uncomplicated, mild TBI cases. Furthermore, the exclusion of extremely severe cases of TBI mitigates the likelihood that the SVTs employed are as robust as implied by the results. As such, independent replication is necessary in these populations as well as in non-traditional neurology samples such as psychiatric samples suffering from cogniform disorder or chronic pain.

The size of the sample also restricted the generalizability of our results. Specifically, the base rate of malingered neurocognitive deficits has been repeatedly shown to occur in 30 – 40% of medico-legal settings. If we had been able to construct our sample size so that it approximated this base rate, the obtained statistical analyses would have better mirrored real-world, clinical settings, therefore increasing the generalizability of the results.

Although the battery was constructed from an assortment of both stand-alone and embedded measures, it may be faulted for its use of only those SVTs that tap memory (as opposed to attention, speed, etc.). Although the literature suggests that the cognitive domain of memory is a highly susceptible to malingering and that feigned memory performance is a

common tactic employed by malingers, it is also probable that a participant might attempt to demonstrate deficits across an assortment of abilities such as other cognitive domains (e.g., attention, motivation, speech), motor coordination, processing speed, and externalizing behaviors. Future research would improve on this study by incorporating a more multidimensional SVT battery, recording behavioral observations, or record the length of time it takes to complete various tasks.

APPENDIX A

Table 1. *Descriptive Statistics Comparing Traumatic Brain Injury (TBI) and Simulator (SIM) Groups*

	TBI (n = 57)		SIM (n = 60)		Total (N = 117)		
Variable	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*Range*
Age (years)	44.6	(11.9)	44.0	(11.4)	43.8	(11.7)	18 – 65
Education (years)	12.2	(2.1)	12.7	(2.0)	12.4	(2.2)	9 – 21
Glasgow Coma Scale	9.4	(3.9)	NA	NA	NA	NA	3 – 15
Time since injury (months)	112.2	(73.9)	NA	NA	NA	NA	10 – 234
Estimated IQ (WTAR)	85.2	(9.1)	92.9	(12.5)	89.3	(11.6)	70 – 122
TOMM Trial 1	44.3	(5.1)	36.7	(11.0)	40.4	(9.2)	9 – 50
TOMM Trial 2	48.4	(3.3)	39.4	(12.8)	43.9	(10.3)	10 – 47
TOMM (% failed)		10.5%		48.3%		29.9%	
Reliable Digit Span (RDS)	7.8	(2.0)	6.7	(2.7)	7.2	(2.5)	0 – 12
RDS (% failed)		24.6%		33.3%		29.1%	
Word Choice Test (WCT)	44.8	(5.9)	39.6	(11.6)	42.2	(9.4)	1 – 50
WCT (% failed)		16.1%		40.0%		28.4%	
MSVT Immediate Recall	96.4	(6.1)	78.9	(25.5)	87.6	(20.3)	5 – 100
MSVT Delayed Recall	94.2	(12.2)	77.2	(26.1)	85.6	21.8	10 – 100
MSVT Consistency	92.8	(13.3)	79.8	(20.9)	86.4	(18.5)	20 – 100
MSVT (% failed)		18.5%		51.7%		35.7%	
CVLT Forced Choice Hits	15.6	(1.0)	14.1	(2.9)	14.8	(2.4)	3 – 16
CVLT (% failed)		7.5%		36.7%		23.0%	

Note. WTAR = Wechsler Test of Adult Reading, Predicted Full Scale IQ, TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CVLT = California Verbal Learning Test-2.

Table 2a. *Descriptive Spearman Correlations for Effort Indices: Simulators (n = 60).*

	1	2	3	4	5	6	7	8	9	10	11	12
1. TOMM Trial 1	1.00											
2. TOMM Trial 2	.89	1.00										
3. Reliable Digits	.32	.40	1.00									
4. Word Choice Test	.68	.76	.44	1.00								
5. MSVT Immediate Recall	.78	.84	.34	.74	1.00							
6. MSVT Delayed Recall	.82	.87	.40	.79	.91	1.00						
7. MSVT-CNS	.72	.81	.28	.69	.86	.90	1.00					
8. CVLT Forced Choice Hits	.60	.70	.62	.71	.61	.68	.58	1.00				
9. Number Failed	-.76	-.83	-.62	-.84	-.85	-.88	-.80	-.83	1.00			
10. Age	.16	.22	.03	.19	.12	.19	.17	.22	-.18	1.00		
11. Education	.00	.03	-.05	-.03	-.06	-.01	.07	-.06	.08	.35	1.00	
12. WTAR Predicted FSIQ	.05	.08	.19	.06	.04	.16	.08	.07	-.03	.18	.46	1.00

Note. TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CNS = Consistency, CVLT = California Verbal Learning Test-2.

*$p < .05$, **$p < .01$.

Table 2b. *Descriptive Spearman Correlations for Effort Indices: TBI (n = 57).*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. TOMM Trial 1	1.00													
2. TOMM Trial 2	.74	1.00												
3. Reliable Digits	.29	.13	1.00											
4. Word Choice Test	.30	.41	.12	1.00										
5. MSVT Immediate Recall	.29	.33	.11	.54	1.00									
6. MSVT Delayed Recall	.34	.40	.26	.54	.59	1.00								
7. MSVT-CNS	.27	.29	.11	.52	.79	.83	1.00							
8. CVLT Forced Choice Hits	.29	.35	.18	.24	.12	.16	.12	1.00						
9. Number Failed	-.36	-.43	-.59	-.58	-.48	-.68	-.53	-.49	1.00					
10. Age	.16	.09	-.24	-.18	-.15	-.13	-.18	-.12	.30	1.00				
11. Education	-.06	-.05	.21	-.16	-.12	-.14	-.16	.03	-.01	.01	1.00			
12. WTAR Predicted IQ	.06	.01	.44	.04	.02	.13	.07	-.05	-.16	.07	.60	1.00		
13. Injury Severity	.08	-.01	-.08	.04	-.04	.07	-.02	-.10	-.01	.19	-.18	-.21	1.00	
14. Months since injury	.04	.05	-.36	-.18	.01	-.10	.03	-.12	.36	.36	-.26	-.27	-.11	1.00

Note. TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CNS = Consistency, CVLT = California Verbal Learning Test-2.

*$p < .05$, **$p < .01$

Table 2c. *Descriptive Spearman Correlations for Effort Indices: Total Sample (N = 117).*

	1	2	3	4	5	6	7	8	9	10	11	12
1. TOMM Trial 1	1.00											
2. TOMM Trial 2	.84	1.00										
3. Reliable Digits	.33	.31	1.00									
4. Word Choice Test	.54	.63	.32	1.00								
5. MSVT Immediate Recall	.65	.71	.29	.67	1.00							
6. MSVT Delayed Recall	.69	.76	.37	.70	.82	1.00						
7. MSVT-CNS	.64	.72	.28	.66	.87	.92	1.00					
8. CVLT Forced Choice Hits	.52	.63	.47	.58	.51	.56	.49	1.00				
9. Number Failed	-.66	-.73	-.60	-.75	-.76	-.83	-.76	-.73	1.00			
10. Age	.15	.15	-.11	.02	.03	.06	.02	.07	.02	1.00		
11. Education	-.08	-.07	.04	-.13	-.14	-.14	-.12	-.09	.11	.16	1.00	
12. WTAR Predicted FSIQ	-.09	-.13	.23	-.04	-.13	-.01	-.06	-.12	.06	.11	.53	1.00

Note. TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CNS = Consistency, CVLT = California Verbal Learning Test-2.

*$p < .05$, **$p < .01$.

Table 3a. *Classification Statistics Based on Published Cutting Scores for Single, Two-, Three-, Four-, and Five-variable Models Predicting Effort Group: TBI* (*n* = 57) *and Simulator* (*n* = 60).

	Hit Rate	Sn	Sp	*Nagelkerke* R^2	AUC	AUC 95% CI	H-L X^2	H-L *p*	BIC
One-Variable Models:									
TOMM	.68	.48	.90	.22	.69	[.59, .79]		--[1]	-406.83
RDS	.54	.33	.75	.01	.54	[.44, .65]		--	-386.63
WCT	.61	.40	.84	.09	.62	[.52, .72]		--	-395.38
MSVT	.66	.52	.82	.16	.67	[.57, .77]		--	-406.43
CVLT	.63	.34	.93	.16	.65	[.54, .75]		--	-406.11
Two-Variable Models:									
TOMM*RDS	.68	.48	.89	.22	.70	[.60, .79]	0.04	.98	-402.22
TOMM*WCT	.68	.48	.89	.22	.69	[.59, .79]	1.06	.59	-403.18
TOMM*MSVT	.69	.50	.89	.24	.70	[.60, .80]	1.19	.55	-409.50
TOMM*CVLT	.69	.53	.87	.28	.72	[.63, .81]	0.72	.70	-413.68
RDS*WCT	.61	.40	.84	.10	.63	[.53, .73]	0.21	.90	-390.86
RDS*MSVT	.66	.52	.82	.16	.67	[.57, .77]	0.66	.72	-401.89
RDS*CVLT	.63	.37	.93	.16	.67	[.56, .76]	2.04	.36	-401.36
WCT*MSVT	.66	.52	.81	.16	.68	[.58, .78]	0.01	1.00	-403.30
WCT*CVLT	.63	.50	.79	.18	.67	[.57, .77]	1.52	.47	-404.82
MSVT*CVLT	.65	.55	.77	.22	.69	[.59, .79]	4.13	.13	-411.77

Note. Sn = Sensitivity (detection of suboptimal effort), Sp = Specificity (bona fide TBI), AUC = Area under the curve, BIC = Bayesian Information Criterion. TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CVLT = California Verbal Learning Test-2 Forced-choice hits.

1. Hosmer-Lemeshow (H-L) statistic not calculated for models with a single dichotomous predictor.

(Table continues…)

Table 3a (continued)

	Hit Rate	Sn	Sp	*Nagelkerke* R^2	AUC	AUC 95% CI	H-L X^2	H-L *p*	BIC
Three-Variable Models:									
TOMM*RDS*WCT	.68	.48	.89	.22	.69	[.59, .79]	1.49	.83	-398.52
TOMM*RDS*MSVT	.71	.55	.87	.24	.70	[.60, .80]	1.91	.75	-404.96
TOMM*RDS*CVLT	.69	.53	.87	.28	.71	[.61, .81]	1.82	.61	-408.92
TOMM*WCT*MSVT	.69	.50	.89	.23	.70	[.61, .80]	1.75	.63	-405.66
TOMM*WCT*CVLT	.69	.53	.87	.28	.73	[.63, .82]	0.99	.80	-409.86
TOMM*MSVT*CVLT	.70	.55	.86	.30	.72	[.63, .82]	1.22	.54	-414.77
RDS*WCT*MSVT	.66	.52	.81	.16	.68	[.58, .78]	0.57	.90	-398.70
RDS*WCT*CVLT	.63	.50	.79	.18	.68	[.59, .78]	2.63	.62	-400.08
RDS*MSVT*CVLT	.65	.55	.77	.22	.68	[.58, .78]	4.39	.36	-407.01
WCT*MSVT*CVLT	.65	.55	.76	.22	.70	[.60, .79]	1.68	.64	-408.32
Four-Variable Models:									
TOMM*RDS*WCT*MSVT	.70	.55	.87	.23	.70	[.60, .80]	1.65	.90	-401.06
TOMM*RDS*WCT*CVLT	.69	.53	.87	.28	.73	[.64, .83]	1.85	.87	-405.09
TOMM*RDS*MSVT*CVLT	.70	.55	.86	.30	.72	[.62, .81]	2.72	.74	-410.06
TOMM*WCT*MSVT*CVLT	.69	.55	.86	.29	.72	[.63, .82]	4.32	.37	-410.91
RDS*WCT*MSVT*CVLT	.65	.55	.76	.22	.69	[.59, .79]	3.64	.46	-403.55
Five-Variable Model									
TOMM*RDS*WCT*MSVT*CVLT	.69	.55	.86	.29	.72	[.62, .82]	2.59	.86	-406.20

Note. Sn = Sensitivity, Sp = Specificity, AUC = Area under the curve, H-L = Hosmer-Lemeshow, BIC = Bayesian Information Criterion, TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CVLT = California Verbal Learning Test-2 Forced-choice hits.

Table 3b. *Logistic Regressions Statistics Based on Published Cutting Scores Predicting Traumatic Brain Injury (TBI) and Simulator (SIM) Group Membership.*

	df	X^2	*p*	*Predictors p < .05*	*Odds Ratio*	*Odds Ratio 95% CI*
One-Variable Models:						
TOMM	1	21.30	< .001	TOMM	7.95	[2.97, 21.31]
RDS	1	1.10	.295	NA	1.54	[0.68, 3.44]
WCT	1	8.40	.004	WCT	3.48	[1.44, 8.40]
MSVT	1	13.91	< .001	MSVT	4.71	[2.00, 11.12]
CVLT	1	14.68	< .001	CVLT	7.09	[2.25, 22.32]
Two-Variable Models:						
TOMM*RDS	2	21.45	< .001	TOMM	7.76	[2.88, 20.93]
TOMM*WCT	2	20.96	< .001	TOMM	6.89	[2.17, 21.90]
TOMM*MSVT	2	21.73	< .001	TOMM	5.66	[1.60, 20.04]
TOMM*CVLT	2	27.00	< .001	TOMM	7.56	[2.18, 26.22]
RDS*WCT	2	8.64	.013	WCT	3.36	[1.38, 8.19]
RDS*MSVT	2	14.13	.001	MSVT	4.53	[1.89, 10.84]
RDS*CVLT	2	14.68	.001	CVLT	7.17	[2.19, 23.46]
WCT*MSVT	2	14.07	.001	MSVT	3.57	[1.25, 10.19]
WCT*CVLT	2	16.62	< .001	CVLT	5.04	[1.50, 16.94]
MSVT*CVLT	2	19.54	< .001	MSVT, CVLT	2.97 4.05	[1.11, 7.94] [1.15, 14.28]

(Table continues…)

Table 3b (continued)

Three-Variable Models:						
TOMM*RDS*WCT	3	21.06	< .001	TOMM	6.81	[2.15, 21.65]
TOMM*RDS*MSVT	3	21.95	< .001	TOMM	5.66	[1.60, 21.08]
TOMM*RDS*CVLT	3	27.01	< .001	TOMM	7.56	[2.18, 26.25]
TOMM*WCT*MSVT	3	21.18	< .001	TOMM	5.51	[1.49, 20.38]
TOMM*WCT*CVLT	3	26.42	< .001	TOMM	7.67	[1.91, 30.76]
TOMM*MSVT*CVLT	3	27.29	< .001	TOMM	7.45	[1.67, 33.24]
RDS*WCT*MSVT	3	14.23	.003	MSVT	3.47	[1.21, 9.97]
RDS*WCT*CVLT	3	16.65	.001	CVLT	5.17	[1.48, 18.03]
RDS*MSVT*CVLT	3	19.54	< .001	MSVT, CVLT	2.96 4.02	[1.11, 7.94] [1.10, 14.65]
WCT*MSVT*CVLT	3	19.32	< .001	CVLT	3.81	[1.07, 13.58]
Four-Variable Models:						
TOMM*RDS*WCT*MSVT	4	21.36	< .001	TOMM	5.52	[1.49, 20.43]
TOMM*RDS*WCT*CVLT	4	26.43	< .001	TOMM	7.66	[1.91, 30.76]
TOMM*RDS*MSVT*CVLT	4	27.36	< .001	TOMM	7.55	[1.68, 33.90]
TOMM*WCT*MSVT*CVLT	4	26.68	< .001	TOMM	7.53	[1.61, 35.35]
RDS*WCT*MSVT*CVLT	4	19.32	.001	CVLT	3.81	[1.03, 14.05]
Five-Variable Model						
TOMM*RDS*WCT*MSVT*CVLT	5	26.72	< .001	TOMM	7.62	[1.62, 35.81]

Note. TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CNS = Consistency, CVLT = California Verbal Learning Test-2 Force-Choice Hits.

Table 4a. *Classification Statistics for Single, Two-, Three-, Four-, and Five-variable Models Predicting Effort Group: TBI* (*n* = 57) *and Simulator* (*n* = 60).

	Hit Rate	Sn	Sp	*Nagelkerke* R^2	AUC	AUC 95% CI	H-L X^2	H-L *p*	BIC
One-Variable Models:									
TOMM	.68	.53	.83	.30	.74	[.65, .83]	2.44	.655	-415.00
RDS	.56	.62	.51	.07	.60	[.50, .71]	3.45	.750	-391.84
WCT	.60	.48	.72	.10	.62	[.51, .72]	9.00	.252	-394.04
MSVTavg	.66	.53	.82	.24	.72	[.62, .81]	.805	.938	-413.20
CVLT	.66	.52	.81	.18	.68	[.58, .78]	.369	.831	-407.66
Two-Variable Models:									
TOMM*RDS	.68	.52	.86	.30	.75	[.66, .84]	2.32	.970	-410.44
TOMM*WCT	.68	.52	.86	.32	.77	[.69, .85]	13.73	.089	-412.55
TOMM*MSVTavg[1]	.68	.54	.83	.30	.73	[.63, .82]	9.56	.144	-416.94
TOMM*CVLT	.68	.53	.85	.33	.76	[.68, .85]	4.08	.538	-418.44
RDS*WCT	.61	.55	.67	.13	.64	[.54, .74]	7.50	.484	-392.34
RDS*MSVTavg	.64	.49	.80	.25	.73	[.63, .82]	4.00	.858	-409.59
RDS*CVLT	.66	.52	.81	.18	.68	[.58, .78]	7.88	.445	-403.29
WCT*MSVTavg	.68	.54	.83	.25	.72	[.63, .81]	16.15	.040	-409.97
WCT*CVLT	.66	.55	.77	.19	.66	[.56, .76]	8.02	.331	-403.63
MSVTavg*CVLT	.65	.53	.78	.28	.73	[.64, .83]	6.50	.260	-417.25

Note. Sn = Sensitivity (detection of suboptimal effort), Sp = Specificity (bona fide TBI), AUC = Area under the curve, H-L = Hosmer-Lemeshow, BIC = Bayesian Information Criterion. TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVTavg = Medical Symptom Validity Test (score average on three subtests), CVLT = California Verbal Learning Test-2 Forced-choice hits.

(Table continues…)

Table 4a (continued)

	Hit Rate	Sn	Sp	*Nagelkerke* R^2	AUC	AUC 95% CI	H-L X^2	H-L *p*	BIC
Three-Variable Models:									
TOMM*RDS*WCT	.68	.52	.86	.32	.77	[.69, .83]	4.80	.779	-408.18
TOMM*RDS*MSVTavg[1]	.68	.53	.85	.30	.60	[.50, .71]	3.39	.908	-410.82
TOMM*RDS*CVLT	.68	.53	.85	.33	.76	[.67, .85]	3.06	.930	-413.93
TOMM*WCT*MSVTavg[1]	.69	.53	.87	.33	.62	[.51, .72]	21.55	.006	-413.69
TOMM*WCT*CVLT	.69	.53	.87	.35	.78	[.70, .87]	12.17	.144	-415.85
TOMM*MSVTavg*CVLT[1]	.69	.54	.86	.33	.72	[.62, .81]	7.77	.256	-417.81
RDS*WCT*MSVTavg	.67	.54	.82	.26	.72	[.62, .81]	4.88	.770	-406.60
RDS*WCT*CVLT	.65	.52	.79	.19	.67	[.57, .77]	6.52	.590	-399.18
RDS*MSVTavg*CVLT	.65	.51	.80	.29	.73	[.64, .83]	5.09	.748	-412.97
WCT*MSVTavg*CVLT	.68	.56	.82	.31	.71	[.61, .81]	6.17	.628	-414.91
Four-Variable Models:									
TOMM*RDS*WCT*MSVT[1]	.71	.56	.87	.34	.68	[.58, .78]	5.47	.706	-409.56
TOMM*RDS*WCT*CVLT	.68	.55	.85	.35	.79	[.71, .81]	9.12	.332	-411.59
TOMM*RDS*MSVTavg*CVLT[1]	.70	.54	.88	.34	.75	[.66, .84]	3.31	.913	-413.59
TOMM*WCT*MSVTavg*CVLT[1]	.72	.56	.90	.36	.77	[.69, .85]	12.20	.143	-416.24
RDS*WCT*MSVTavg*CVLT	.66	.56	.78	.31	.73	[.63, .82]	6.48	.593	-410.65
Five-Variable Model									
TOMM*RDS*WCT*MSVTavg*CVLT[1]	.71	.56	.88	.37	.78	[.70, .87]	6.66	.574	-412.17

Note. Sn = Sensitivity, Sp = Specificity, AUC = Area under the curve, BIC = Bayesian Information Criterion, TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CVLT = California Verbal Learning Test-2 Forced-choice hits.

1. Models containing both TOMM and MSVTavg may be unstable due to high collinearity.

Table 4b. *Logistic Regression Statistics: Predicting Traumatic Brain Injury (TBI) and Simulator (SIM) Group Membership.*

	df	X^2	*p*	*Predictors p < .05*	*Odds Ratio Exp(B)*	*Odds Ratio 95% CI*
One-Variable Models:						
TOMM	1	29.47	< .001	TOMM	0.85	[0.78, 0.93]
RDS	1	6.30	.012	RDS	0.82	[0.69, 0.97]
WCT	1	9.51	.002	WCT	0.93	[0.89, 0.98]
MSVT	1	21.98	< .001	MSVT	0.94	[0.92, 0.97]
CVLT	1	16.22	< .001	CVLT	0.60	[0.43, 0.84]
Two-Variable Models:						
TOMM*RDS	2	29.67	< .001	TOMM	0.86	[0.79, 0.93]
TOMM*WCT	2	31.78	< .001	TOMM	0.81	[0.72, 0.91]
TOMM*MSVT	2	28.66	< .001	TOMM	0.87	[0.77, 0.98]
TOMM*CVLT	2	31.73	< .001	TOMM	0.83	[0.74, 0.94]
RDS*WCT	2	11.57	.003	WCT	0.95	[0.90, 0.99]
RDS*MSVT	2	23.13	< .001	MSVT	0.95	[0.92, 0.98]
RDS*CVLT	2	16.62	< .001	CVLT	0.64	[0.44, 0.89]
WCT*MSVT	2	23.51	< .001	MSVT	0.92	[0.88, 0.97]
WCT*CVLT	2	16.96	< .001	CVLT	0.65	[0.45, 0.94]
MSVT*CVLT	2	26.27	< .001	MSVT	0.95	[0.91, 0.98]

(Table continues…)

Table 4b (continued)

Three-Variable Models:						
TOMM*RDS*WCT	3	32.17	< .001	TOMM	0.81	[0.72, 0.91]
TOMM*RDS*MSVT	3	29.12	< .001	TOMM	0.88	[0.78, 0.99]
TOMM*RDS*CVLT	3	32.02	< .001	TOMM	0.83	[0.74, 0.94]
TOMM*WCT*MSVT	3	32.00	< .001	TOMM	0.85	[0.75, 0.97]
TOMM*WCT*CVLT	3	33.94	< .001	TOMM	0.79	[0.68, 0.92]
TOMM*MSVT*CVLT	3	31.59	< .001	--	0.86	[0.74, 1.00]
RDS*WCT*MSVT	3	24.90	< .001	MSVT	0.93	[0.89, 0.97]
RDS*WCT*CVLT	3	17.27	< .001	CVLT	0.67	[0.45, 0.99]
RDS*MSVT*CVLT	3	26.75	< .001	MSVT	0.95	[0.91, 0.98]
WCT*MSVT*CVLT	3	28.98	< .001	MSVT	0.92	[0.87, 0.98]
Four-Variable Models:						
TOMM*RDS*WCT*MSVT	4	32.63	< .001	TOMM	0.85	[0.75, 0.98]
TOMM*RDS*WCT*CVLT	4	34.45	< .001	TOMM	0.79	[0.68, 0.92]
TOMM*RDS*MSVT*CVLT	4	32.14	< .001	--	0.86	[0.73, 1.00]
TOMM*WCT*MSVT*CVLT	4	34.79	< .001	TOMM	0.84	[0.71, 0.99]
RDS*WCT*MSVT*CVLT	4	29.20	< .001	MSVT	0.92	[0.88, 0.98]
Five-Variable Model						
TOMM*RDS*WCT*MSVT*CVLT	5	35.48	< .001	TOMM	0.83	[0.71, 0.99]

Note. TOMM = Test of Memory Malingering, RDS = Reliable Digit Span, WCT = Word Choice Test, MSVT = Medical Symptom Validity Test, CNS = Consistency, CVLT = California Verbal Learning Test-2 Force-Choice Hits.

Table 5. *Classification Concordance for Pairs of Seven Performance Indices: TBI Group (n = 57), Simulator Group (n = 60), and Total Sample* (*N* = 117).

TOMM * MSVT Classifications: TBI					
			MSVT		
			Pass	Fail	*Total*
TOMM	Pass	Count	42	6	*48*
		% of Total	77.8	11.1	*88.9*
	Fail	Count	2	4	*6*
		% of Total	3.7	7.4	*11.1*
Total		*Count*	*44*	*10*	*54*
		% of Total	*81.5*	*18.5*	*100.0*

TOMM * MSVT Classifications: Simulator					
			MSVT		
			Pass	Fail	*Total*
TOMM	Pass	Count	25	4	*29*
		% of Total	43.1	6.9	*50.0*
	Fail	Count	3	26	*29*
		% of Total	5.2	44.8	*50.0*
Total		*Count*	*28*	*30*	*58*
		% of Total	*48.3*	*51.7*	*100.0*

TOMM * MSVT Classifications: Total					
			MSVT		
			Pass	Fail	*Total*
TOMM	Pass	Count	67	10	*77*
		% of Total	59.8	8.9	*68.8*
	Fail	Count	5	30	*35*
		% of Total	4.5	26.8	*31.3*
Total		*Count*	*72*	*40*	*112*
		% of Total	*64.3*	*35.7*	*100.0*

TOMM * RDS Classifications: TBI					
			RDS		
			Pass	Fail	*Total*
TOMM	Pass	Count	39	12	*51*
		% of Total	68.4	21.1	*89.5*
	Fail	Count	4	2	*6*
		% of Total	7.0	3.5	*10.5*
Total		*Count*	*43*	*14*	*57*
		% of Total	*75.4*	*24.6*	*100.0*

TOMM * RDS Classifications: Simulator					
			RDS		
			Pass	Fail	*Total*
TOMM	Pass	Count	23	8	*31*
		% of Total	38.3	13.3	*51.7*
	Fail	Count	17	12	*29*
		% of Total	28.3	20.0	*48.3*
Total		*Count*	*40*	*20*	*60*
		% of Total	*66.7*	*33.3*	*100.0*

TOMM * RDS Classifications: Total					
			RDS		
			Pass	Fail	*Total*
TOMM	Pass	Count	62	20	*82*
		% of Total	53.0	17.1	*70.1*
	Fail	Count	21	14	*35*
		% of Total	17.9	12.0	*29.9*
Total		*Count*	*83*	*34*	*117*
		% of Total	*70.9*	*29.1*	*100.0*

Table 5 (continued). *Classification Concordance for Pairs of Seven Performance Indices: TBI Group (n = 57), Simulator Group (n = 60), and Total Sample* (*N* = 117).

TOMM * WCT Classifications: TBI					
			WCT		
			Pass	Fail	*Total*
TOMM	Pass	Count	44	6	*50*
		% of Total	78.6	10.7	*89.3*
	Fail	Count	3	3	*6*
		% of Total	5.4	5.4	*10.7*
Total		*Count*	*47*	*9*	*56*
		% of Total	*83.9*	*16.1*	*100.0*

TOMM * WCT Classifications: Simulator					
			WCT		
			Pass	Fail	*Total*
TOMM	Pass	Count	28	3	*31*
		% of Total	46.7	5.0	*51.7*
	Fail	Count	8	21	*29*
		% of Total	13.3	35.0	*48.3*
Total		*Count*	*36*	*24*	*60*
		% of Total	*60.0*	*40.0*	*100.0*

TOMM * WCT Classifications: Total					
			WCT		
			Pass	Fail	*Total*
TOMM	Pass	Count	72	9	*81*
		% of Total	62.1	7.8	*69.8*
	Fail	Count	11	24	*35*
		% of Total	9.5	20.7	*30.2*
Total		*Count*	*83*	*33*	*116*
		% of Total	*71.6*	*28.4*	*100.0*

TOMM * $CVLT_{FC}$ Classifications: TBI					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
TOMM	Pass	Count	46	3	*49*
		% of Total	86.8	5.7	*92.5*
	Fail	Count	3	1	*4*
		% of Total	5.7	1.9	*7.5*
Total		*Count*	*49*	*4*	*53*
		% of Total	*92.5*	*7.5*	*100.0*

TOMM * $CVLT_{FC}$ Classifications: Simulator					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
TOMM	Pass	Count	28	3	*31*
		% of Total	46.7	5.0	*51.7*
	Fail	Count	10	19	*29*
		% of Total	16.7	31.7	*48.3*
Total		*Count*	*38*	*22*	*60*
		% of Total	*63.3*	*36.7*	*100.0*

TOMM * $CVLT_{FC}$ Classifications: Total					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
TOMM	Pass	Count	74	6	*80*
		% of Total	65.5	5.3	*70.8*
	Fail	Count	13	20	*33*
		% of Total	11.5	17.7	*29.2*
Total		*Count*	*87*	*26*	*113*
		% of Total	*77.0*	*23.0*	*100.0*

Table 5 (continued). *Classification Concordance for Pairs of Seven Performance Indices: TBI Group (n = 57), Simulator Group (n = 60), and Total Sample* (*N* = 117).

MSVT * RDS Classifications: TBI					
			RDS		
			Pass	Fail	*Total*
MSVT	Pass	Count	34	10	*44*
		% of Total	63.0	18.5	*81.5*
	Fail	Count	7	3	*10*
		% of Total	13.0	5.6	*18.5*
Total		*Count*	*41*	*13*	*54*
		% of Total	*75.9*	*24.1*	*100.0*

MSVT * RDS Classifications: Simulator					
			RDS		
			Pass	Fail	*Total*
MSVT	Pass	Count	22	6	*28*
		% of Total	37.9	10.3	*48.3*
	Fail	Count	16	14	*30*
		% of Total	27.6	24.1	*51.7*
Total		*Count*	*38*	*20*	*58*
		% of Total	*65.5*	*34.5*	*100.0*

MSVT * RDS Classifications: Total					
			RDS		
			Pass	Fail	*Total*
MSVT	Pass	Count	56	16	*72*
		% of Total	50.5	14.3	*64.3*
	Fail	Count	23	17	*40*
		% of Total	20.5	15.2	*35.7*
Total		*Count*	*79*	*33*	*112*
		% of Total	*70.5*	*29.5*	*100.0*

MSVT * WCT Classifications: TBI					
			WCT		
			Pass	Fail	*Total*
MSVT	Pass	Count	40	3	*43*
		% of Total	75.5	5.7	*81.1*
	Fail	Count	4	6	*10*
		% of Total	7.5	11.3	*18.9*
Total		*Count*	*44*	*9*	*53*
		% of Total	*83.0*	*17.0*	*100.0*

MSVT* WCT Classifications: Simulator					
			WCT		
			Pass	Fail	*Total*
MSVT	Pass	Count	25	3	*28*
		% of Total	43.1	5.2	*48.3*
	Fail	Count	9	21	*30*
		% of Total	15.5	36.2	*51.7*
Total		*Count*	*34*	*24*	*58*
		% of Total	*58.6*	*41.4*	*100.0*

MSVT* WCT Classifications: Total					
			WCT		
			Pass	Fail	*Total*
MSVT	Pass	Count	65	6	*71*
		% of Total	58.6	5.4	*64.0*
	Fail	Count	13	27	*40*
		% of Total	11.7	24.3	*36.0*
Total		*Count*	*78*	*33*	*111*
		% of Total	*70.3*	*29.7*	*100.0*

Table 5 (continued). *Classification Concordance for Pairs of Seven Performance Indices: TBI Group (n = 57), Simulator Group (n = 60), and Total Sample* (*N* = 117).

MSVT * $CVLT_{FC}$ Classifications: TBI					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
MSVT	Pass	Count	39	3	*42*
		% of Total	76.5	5.9	*82.4*
	Fail	Count	8	1	*9*
		% of Total	15.7	2.0	*17.6*
Total		*Count*	*47*	*4*	*51*
		% of Total	*92.2*	*7.8*	*100.0*

MSVT * $CVLT_{FC}$ Classifications: Simulator					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
MSVT	Pass	Count	26	2	*28*
		% of Total	44.8	3.4	*48.3*
	Fail	Count	10	20	*30*
		% of Total	17.2	34.5	*51.7*
Total		*Count*	*36*	*22*	*58*
		% of Total	*62.1*	*37.9*	*100.0*

MSVT * $CVLT_{FC}$ Classifications: Total					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
MSVT	Pass	Count	65	5	*70*
		% of Total	59.6	4.6	*64.2*
	Fail	Count	18	21	*39*
		% of Total	16.5	19.3	*35.8*
Total		*Count*	*83*	*26*	*109*
		% of Total	*76.1*	*23.9*	*100.0*

RDS * WCT Classifications: TBI					
			WCT		
			Pass	Fail	*Total*
RDS	Pass	Count	36	6	*42*
		% of Total	64.3	10.7	*75.0*
	Fail	Count	11	3	*14*
		% of Total	19.6	5.4	*25.0*
Total		*Count*	*47*	*9*	*56*
		% of Total	*83.9*	*16.1*	*100.0*

RDS * WCT Classifications: Simulator					
			WCT		
			Pass	Fail	*Total*
RDS	Pass	Count	27	13	*40*
		% of Total	45.0	21.7	*66.7*
	Fail	Count	9	11	*20*
		% of Total	15.0	18.3	*33.3*
Total		*Count*	*36*	*24*	*60*
		% of Total	*60.0*	*40.0*	*100.0*

RDS * WCT Classifications: Total					
			WCT		
			Pass	Fail	*Total*
RDS	Pass	Count	63	19	*82*
		% of Total	54.3	16.4	*70.7*
	Fail	Count	20	14	*34*
		% of Total	17.2	12.1	*29.3*
Total		*Count*	*83*	*33*	*116*
		% of Total	*71.6*	*28.4*	*100.0*

Table 5 (continued). *Classification Concordance for Pairs of Seven Performance Indices: TBI Group (n = 57), Simulator Group (n = 60), and Total Sample* (*N* = 117).

RDS * $CVLT_{FC}$ Classifications: TBI					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
RDS	Pass	Count	37	3	*40*
		% of Total	69.8	5.7	*75.5*
	Fail	Count	12	1	*13*
		% of Total	22.6	1.9	*24.5*
Total		*Count*	*49*	*4*	*53*
		% of Total	*92.5*	*7.5*	*100.0*

RDS * $CVLT_{FC}$ Classifications: TBI					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
RDS	Pass	Count	31	9	*40*
		% of Total	51.7	15.0	*66.7*
	Fail	Count	7	13	*20*
		% of Total	11.7	21.7	*33.3*
Total		*Count*	*38*	*22*	*60*
		% of Total	*63.3*	*36.7*	*100.0*

RDS * $CVLT_{FC}$ Classifications: Total					
			$CVLT_{FC}$		
			Pass	Fail	*Total*
RDS	Pass	Count	68	12	*80*
		% of Total	60.2	10.6	*70.8*
	Fail	Count	19	14	*33*
		% of Total	16.8	12.4	*29.2*
Total		*Count*	*87*	*26*	*113*
		% of Total	*77.0*	*23.0*	*100.0*

WCT * $CVLT_{FC}$ Classifications: TBI					
			Pass	Fail	*Total*
RDS	Pass	Count	41	3	*44*
		% of Total	78.8	5.8	*84.6*
	Fail	Count	7	1	*8*
		% of Total	13.5	1.9	*15.4*
Total		*Count*	*48*	*4*	*52*
		% of Total	*92.3*	*7.7*	*100.0*

WCT * $CVLT_{FC}$ Classifications: Simulator					
			Pass	Fail	*Total*
RDS	Pass	Count	30	6	*36*
		% of Total	50.0	10.0	*60.0*
	Fail	Count	8	16	*24*
		% of Total	13.3	26.7	*40.0*
Total		*Count*	*38*	*22*	*60*
		% of Total	*63.3*	*36.7*	*100.0*

WCT * $CVLT_{FC}$ Classifications: Total					
			Pass	Fail	*Total*
RDS	Pass	Count	71	9	*80*
		% of Total	63.4	8.0	*71.4*
	Fail	Count	15	17	*32*
		% of Total	13.4	15.2	*28.6*
Total		*Count*	*86*	*26*	*112*
		% of Total	*76.8*	*23.2*	*100.0*

REFERENCES

Binder, L. M., & Kelly, M. P. (1994). Portland Digit Recognition test (PDRT) performance by motivationally intact patients with brain dysfunction. *Archives of Clinical Neuropsychology, 9(2),* 111.

Bush, S. S., Ruff, R. M., Tröster, 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 and 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. *The Clinical Neuropsychologist, 22(3)*, 440.

Cohen, J. (1960). A coefficient of agreement for nominal scale. *Educational and Psychological Measurement, 20(1),* 37-46.

Coleman, R. D., Rapport, L. J., Millis, S. R., Ricker, J. H., & Falchion, 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.

Constantinou, M., Bauer, L., Ashendorf, L., Fisher, J.M., & McCaffrey, R.J. (2005). Is poor performance on recognition memory effort measures indicative of generalized poor performance on neuropsychological tests? *Archives of Clinical Neuropsychology, 20(2),* 191-198.

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. *The 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.

Faul, M., Xu, L., Wald, M. M., & Coronado, V. G. (2010). Traumatic brain injury in the United States: Emergency Department visits, hospitalizations, and deaths 2002 - 2006. Atlanta, GA: Centers for Disease Control and Prevention: National Center for Injury Prevention and Control.

Finkelstein, E., Coors, P., & Miller, T. (2006). *The Incidence and Economic Burden of Injuries in the United States.* New York (NY): Oxford University Press.

Frederick, R., & Bowden, S. (2009). Evaluating constructs represented by symptom validity tests in forensic neuropsychological assessment of traumatic brain injury. *Journal of Head Trauma Rehabilitation. 24(2),* 105-122.

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

Green, P. (2005). *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. *The Clinical Neuropsychologist, 9(3),* 230-240.

Greiffenstein, M. F., Greve, K. W., Bianchini, K. J., & Baker, W. J. (2008). Test of Memory

Malingering and Word Memory Test: A new comparison of failure concordance rates. *Archives of Clinical Neuropsychology 23(7-8),* 801–807.

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. *The Clinical Neuropsychologist, 24(1),* 137-152.

Harrell, F. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer-Verlag.

Harrell, F. E. Jr., Lee, K. L., Califf, R. M., Pryor, D. B., & Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine, 3(2),* 143-152.

Harrell, F. E. Jr., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15(4),* 361-387.

Harrell, F.E. Jr., Lee, K.L., Matchar, D.B., & Reichert, T.A. (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treatment Reports, 69(10),* 1071-1077.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23(7),* 1093–1129.

Hiscock, M., & Hiscock, C. K. (1989). Refining the Forced-Choice Method for the Detection of Malingering. *Journal of Clinical and Experimental Neuropsychology, 11(6),* 967-74.

Hosmer, D. W., & Lemeshow, S. (2000). Applied logisitic regression. New York: Wiley

Interscience.

Inman, T. H., & Berry, D.T.R. (2002). Cross-validation of indicators of malingering - A comparison of nine neuropsychological tests, four tests of malingering, and behavioral observations. *Archives of Clinical Neuropsychology 17(1),* 1-23.

Johnstone, B., Hexum, C., & Ashkanazi, G. (1995). Extent of cognitive decline in traumatic brain injury based on estimates of premorbid intelligence. *Brain Injury. 9(4),* 377-384.

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

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

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, J., & Fischer, J. (2004). Neuropsychological Assessment. Fourth Edition. Oxford University Press.

Lynch, W. J. (2004). Determination of effort level, exaggeration, and malingering in neurocognitive assessment. *Journal of Head Trauma Rehabilitation, 19(3),* 277–283.

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

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

Millis, S. R. (1992). The Recognition Memory Test in the detection of malingered and exaggerated memory deficits. *The Clinical Neuropsychologist, 6,* 404-414.

Millis, S. R., & Putnam, S. J. (1994). The Recognition Memory Test in the assessment of memory impairment after financially compensable mild head injury: A replication. *Perceptual and Motor Skills, 79(1),* 384–386.

Millis, S. R. (2010). What clinicians really need to know about symptom exaggeration, insufficient effort, and malingering: Statistical and measurement matters. In J. E. Morgan, I. S. Baron, & J. H. Ricker (Eds.), *Casebook of Clinical Neuropsychology.* Oxford University Press, USA.

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.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika. 78(3),* 691-692.

Pankratz, L., & Binder, L. M. (1997). Malingering on intellectual and neuropsychological measures. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 223–236). New York: Guilford Press.

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

Raferty, A. (1995). Bayesian Model Selection in Social Research. *Social Methodology, 25,* 111-163, Oxford, UK: Basil Blackwell.

Rapport, L. J., Farchione, T. J., Coleman, R. D., & Axelrod, B. N. (1998). Effects of coaching on malingered motor function profiles. *Journal of Clinical Experimental Neuropsychology, 20(1),* 89-97.

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.

Rees, L. M., T. N. Tombaugh, et al. (2001). Depression and the Test of Memory Malingering. *Archives of Clinical Neuropsychology 16(5),* 501-506.

Richman, J., Green, P., Gervais, R., Flare, L., Marten, T., Brockhaus, R., et al. (2006). Objective tests of symptom exaggeration in independent medical examinations. *Journal of Occupational and Environmental Medicine, 48(3),* 303-311.

Richman, J., P. Green, et al. (2006). Objective tests of symptom exaggeration in independent medical examinations. *Journal of Occupational and Environmental Medicine 48(3),* 303-311.

Riggo, S. (2011). Traumatic brain injury and its neurobehavioral sequelae. *Neurologic Clinics, 29(1)*, 35-47.

Rogers, R. (Ed.). (1988). *Clinical Assessment of malingering and deception.* New York: Guilford Press.

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

Rosenfeld, B., Sands, S., & Van Gorp, W. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology, 15(4),* 349-359.

Seel, R., Kreutzer, J., et al. (2003). Depression after traumatic brain injury: A National Institute on Disability and Rehabilitation Research Model Systems multicenter investigation. *Archives of Physical Medicine and Rehabilitation, 84(2)*, 177-184.

Sharland, M. J. & Gfeller, J. D. (2007). A survey of neuropsychologists' beliefs and practices with respect to the assessment of effort. *Archives of Clinical Neuropsychology 22(2),* 213-223.

Slick, D. J., Tan, J. E., et al. (2004). Detecting malingering: a survey of experts' practices. *Archives of Clinical Neuropsychology 19(4),* 465-473.

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

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

Slick, D. J., Hopp, G., Strauss, E., & Spellacy, F. J. (1996). Victoria Symptom Validity Test: Efficiency for detecting feigned memory impairment and relationship to neuropsychological tests and MMPI-2 validity scales. *Journal of Clinical and Experimental Neuropsychology, 18(6),* 911–922.

Strauss, E., Sherman, E., & Spreen, O. (2006). A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary (3rd Ed.). Oxford University Press.

Tan, J. E., D. J. Slick, et al. (2002). How'd they do it? Malingering strategies on symptom validity tests. *The Clinical Neuropsychologist 16(4),* 495-505.

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.

The Psychological Corporation. (2001). Wechsler Test of Adult Reading. San Antonio, TX:

Harcourt Brace & Company.

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

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

Vallabhajosula, B., & Van Gorp, W. G. (2001). Post-Daubert admissibility of scientific evidence on malingering of cognitive deficits. *Journal of the American Academy of Psychiatry and the Law, 29(2),* 207–215.

Victor, T., Boone, K., et al. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist. 23(2),* 297-313.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale (4th ed.).* San Antonio, TX: Psychological Corporation.

Wolfe, P. L., Millis, S. R., Hanks, R., Fichtenberg, N., Larrabee, G. J., Sweet, J. (2010). Effort indicators within the California Verbal Learning Test-II (CVLT-II). *The Clinical Neuropsychologist, 24(1),* 153-68.

ABSTRACT

**DETECTING SUBOPTIMAL EFFORT
IN TRAUMATIC BRAIN INJURY ASSESSMENT**

by

JESSE R. BASHEM

August 2012

Advisor: Dr. Lisa J. Rapport

Major: Psychology

Degree: Master of Arts

Purposeful presentation of suboptimal effort is a primary pitfall to accurate assessment, especially among individuals seeking compensation. It is known that successful simulation of impairment becomes increasingly difficult when feigning is required across multiple measures. This study evaluated the diagnostic efficiencies and predictive ability of five symptom validity tests: Test of Memory Malingering (TOMM), Medical Symptom Validity Test (MSVT), California Verbal Learning Test – Forced Choice (CVLT), Reliable Digit Span, and Word Choice Test. Participants were 57 adults with moderate to severe traumatic brain injury and 60 healthy adults coached to simulate memory impairment. Tests were evaluated using logistic regression, ROC curve, and Bayesian Information Criterion statistics. Results indicate that the TOMM and MSVT performed best; however, they operated less effectively than combined use of the TOMM and CVLT in differentiating bona fide TBI and simulators. The limitations of comparing multivariable models psychometrically are discussed, as are areas of future research.

AUTOBIOGRAPHICAL STATEMENT

Education

June 2005	**Bachelor of Arts** University of California, Santa Cruz, California **Major:** Psychology & Philosophy

Clinical Experience

September 2011 – Present	**Wayne State University Counseling and Psychological Services** Provided WSU students with individual, couples, or group psychotherapy, assessments, psychoeducational workshops, and outreach presentations. Advisor: Kristin Van de Laar, Ph.D.
September 2010 – August 2011	**Center for Forensic Psychiatry** Psychological assessment and psychotherapy for inpatients found incompetent to stand trial or not guilty by reason of insanity. Advisor: Judith Shazer, Ph.D.
March 2009 – April 2010	**Rehabilitation Institute of Michigan** Neuropsychological Assessment of Traumatic Brain Injury
August 2008 – Present	**Wayne State University Psychology Clinic** Individual Psychological Assessment & Psychotherapy Dialectical Behavioral Therapy group co-leader

Research Experience

March 2009 – Present	**Collaborative Investigator, Thesis Mentee** Lisa Rapport, Ph.D., Professor, Clinical Psychology; Wayne State University, Detroit, Michigan
December 2004 – May 2005	**Research Assistant** Developmental Psychology University of California, Santa Cruz, California Advisor: Catherine Cooper, Ph.D., Professor

Competitive Funding & Awards:

2011	APA Invitation to the 5th Annual Psychological Science Graduate Student Superstars – Datablitz
2011	APA Division 40 Applied Neuropsychology Student Poster Award
2010	American Psychological Association Student Travel Award
2008 – 2009	Thomas C. Rumble Fellowship – Wayne State University