2012-07-03

Malingering Detection Measure Utility and Concordance in a University Accommodation-Seeking Student Population

Nichole M. Loser
Brigham Young University - Provo

Malingering Detection Measure Utility and Concordance in a University Accommodation-Seeking Student Population

Nichole M. Loser

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Bruce N. Carpenter, Chair
Sally Barlow
Scott Braithwaite
Michael Larson
Patrick Steffan

Department of Clinical Psychology

Brigham Young University

August 2012

ABSTRACT

Malingering Detection Measure Utility and Concordance in a
University Accommodation-Seeking Student Population

Nichole M. Loser
Department of Clinical Psychology, BYU
Doctor of Philosophy

According to the Americans with Disabilities Act, universities and colleges are required to provide accommodative services for students with disabilities. Many studies have examined the role of malingering mental health symptoms in order to obtain psychotropic medications, but very little research has been done on the role of accommodations as secondary gain in students who may malinger learning disabilities. This study sought to examine both the usefulness of implementing specific malingering detection measures in psychological evaluations with university students and the agreement of those measures within the population. Archival data was gathered from a university accommodation clinic that provided free psychological evaluations for consecutively presenting students (N=121). Four malingering detection measures were used: the Test of Memory and Malingering (TOMM), the Word Memory Test (WMT), the WAIS Digit Span (DS) and two cut scores for the MMPI-2 F Scale (F Scale 80 and F Scale 95). Scores for these four malingering detection measures were compared in terms of their agreement rates, their classification rates (at a 10% malingering base rate recommendation), and their sensitivity, specificity, positive and negative predictive powers using both the TOMM and WMT independently as diagnostic criterion. A qualitative examination of the data revealed that different combinations of measures did classify some of the same respondents as malingering. Results indicated that each of these four measures share the ability to detect malingering in its different forms and have similar classification rates. Although the TOMM and WMT likely provide overlapping information, the pragmatic implementation of one of these measures may assist in the evaluation of suspected malingering with accommodation-seeking students.

keywords: malingering, accommodation, student population, TOMM, WMT, F Scale, Digit Span

ACKNOWLEDGEMENTS

I would like to thank my committee, and especially my chair, Dr. Bruce Carpenter, for their assistance and aid in completing this final step of my graduate school experience. I would also like to thank Dr. Michael Brooks, Director of the University Accessibility Center, for facilitating access to the foundational data of my dissertation, for excellent initial guidance and direction, and for meetings that included references to South Park. I would like to thank my family and friends for their support during my undertaking of a chaotic graduate school and dissertation life. Lastly, I would like to thank my father, Richard. I hope you are well pleased with me.

TABLE OF CONTENTS

Abstract ii

Acknowledgements iii

Table of Contents iv

List of Tables vi

Introduction 1

Malingering in an Accommodation-Seeking Student Population 3

What is Malingering? 5

Malingering Detection Measures 6

Diagnostic considerations 7

Symptom validity tests 8

Effort tests 8

Four Malingering Detection Measures 11

The Test of Memory and Malingering 11

The Word Memory Test 12

The MMPI-2 F-Scale 12

WAIS Digit Span 13

Malingering Detection Measure Concordance 14

Present Study Hypotheses 15

Method 17

Procedures 17

Participants 18

Measures 18

The Test of Memory and Malingering 18
The Word Memory Test 19
The MMPI-2 F-Scale 20
WAIS Digit Span 20
Results 21
Comparison of Measure Classification Rates.....21
Psychometric Properties and Inter-test Correlations 26
Qualitative Comparisons 29
Discussion 30
Limitations 34
Summary and Future Directions 36
References 37

LIST OF TABLES

Table 1	Demographic Characteristics	22
Table 2	Classification Rates of Malingering	23
Table 3	Agreement Rates between TOMM1 and [WMTIR, F Scale 95, F Scale 80, and DS]	24
Table 4	Agreement rates between WMTIR and [F Scale 95, F Scale 80, and DS]	25
Table 5	Agreement rates between [F Scale 95 and DS] and [F Scale 80 and DS]	25
Table 6	Cut Score differences between the F Scale 95 and F Scale 80	26
Table 7	Sensitivity, Specificity, Positive Predictive Value and Negative Predictive Value of Four Malingering Measures using the TOMM1	26
Table 8	Sensitivity, Specificity, Positive Predictive Value and Negative Predictive Value of Four Malingering Measures using the WMTIR	27
Table 9	Intercorrelations among Malingering Detection Measures	29

Malingering Detection Measure Utility and Concordance in a University Accommodation-Seeking Student Population

According to the Americans with Disabilities Act, education-seeking individuals with mental health disabilities, including attention and learning disorders, have a legal right to access specialized resources in educational settings (Petrila & Brink, 2001). In recent decades, this disability legislation has been made compulsory among educational institutions (Latham, 2005), forcing colleges and universities to address the issue of dwindling resources coupled with ever-increasing demand for accommodations. This expedites the need to allocate both monetary and staffing resources to students who do indeed have learning disabilities and who are legally entitled to such accommodations. Unfortunately, recent research indicates there is a growing movement among higher education students to seek out school resources they do not actually need or qualify for, placing additional strain on schools.

These specialized resources may include psychiatric consultation and medication, additional time on individual class assignments and tests, elimination of spelling and grammar penalties, personal note takers, private testing rooms, reduced homework, textbooks on tape, alternate test forms, special seating in distraction-free testing rooms, and additional time on college and graduate program entrance exams (McGuire, 1998). These accommodations are not meant to provide disabled individuals with an advantage in their academic progress. Rather, they are intended to level the proverbial playing field, because they are based on the assumption that learning-disabled students are at a disadvantage in their ability to learn and perform as well as non-disabled students in a typical university environment.

Due to limited available funding, the majority of higher education institutions have limited resources to offer these accommodation-seeking students. Thus, in order to provide these services, colleges and universities often require documentation of an individual's learning disability. This documentation has historically been gleaned from measures of a student's innate potential and their academic progress via a psychological evaluation. Many students also receive diagnoses from psychiatric providers, and those with diagnoses typically qualify for disability-specific services, based on availability, at their college or university.

In light of these circumstances, recent studies have focused on determining if students in accommodation-seeking populations fake or distort their symptoms or deficits for personal gain. This manipulation of symptoms is commonly known as malingering. There are a large number of studies addressing malingering in populations with Attention-Deficit/Hyperactivity Disorder (ADHD) diagnoses (McCabe, Teter, & Boyd, 2006; White, Becker-Blease, & Grace-Bishop, 2006), and there is a plethora of research available detailing the unnecessary acquisition of prescription stimulant medications for fabricated attention and concentration difficulties, for both illegal personal use and for profit (Advokat, Guidry, & Martino, 2008; Harrison, 2006; Harrison, Edwards, & Parker, 2007; Jachimowicz & Geiselman, 2004; McCann & Roy-Byrne, 2004; Quinn, 2003; Suhr, Hammers, Dobbins-Buckland, Zimak, & Hughes, 2008).

Researchers have proposed that if disability accommodations provide enough incentive for students to malinger ADHD diagnoses, then an extension to learning disorder malingering is highly probable. As predicted, new studies conducted on learning disorder assessments in this same population indicate similar attempts at inappropriate resource acquisition (Frazier, Frazier, Busch, Kerwood, & Demaree, 2008; Harrison & Edwards, 2010; Sullivan, May, & Galbally, 2007). As college students seek out resources they do not qualify for, it may be prudent to

evaluate the usefulness of introducing malingering detection measures (MDMs) as part of traditional learning disability evaluations. In addition, due to the large number of measures available, examination of test concordance or agreement in this population may assist clinicians in determining which measures should be introduced into the assessment battery.

This study attempted to address both the utility of MDMs and their concordance in this population by comparing four malingering measures across domains with several primary goals. First, it determined if each MDM detected malingering relative to predicted base rates. Second, it examined the concordance or agreement rates for each MDM to establish if these measures classified the same individuals as malingering. Third, it compared the performance of embedded MDMs with measures that are dedicated to the detection of malingering, as a comparison of detection rates between these two groups may be useful in helping university accommodation coordinators determine the utility of administering additional tests, or if embedded measures are sufficient to screen for potential malingering. As a whole, this information may be very useful in predicting if these four MDMs provide any utility or usefulness in a secondary education accommodation-seeking population and if so, which measures may be most useful according to financial, staffing, and time constraints.

Malingering in an Accommodation-Seeking Student Population

In spite of research indicating there is a high likelihood for learning-disordered malingering at higher education institutions (Blanchard, McGrath, Pogge, & Khadivi, 2003; Crank & Deshler, 2001; Lindstrom, Lindstrom, Coleman, Nelson, & Gregg, 2009), there are very few studies that actually address potential exaggeration with the use of MDMs. The few studies that have been done indicate that base rates of malingering for this population are between 8-12% at two universities in the United States (Clayton, 2011; Harrison et al., 2007) and at 14.6%

for a university in Canada (Harrison & Edwards, 2010). Higher education institutions document learning disabilities through a battery of psychological tests that, until recently, has not included the assessment of potential symptom exaggeration or malingering. This may be an important oversight, as base rates for malingering indicate that colleges and universities may be spending more than 10% of their accommodation resources on non-deserving students.

Although traditional MDMs have been well-studied and validated in neuropsychological testing batteries (Bauer, O'Bryant, Lynch, McCaffrey, & Fisher, 2007; Constantinou, Bauer, Ashendorf, Fisher, & McCaffrey, 2005; Green, Lees-Haley, & Allen, 2002; Larrabee, 2003b; Rogers, Harrell, & Liff, 1993; Weinborn, Orr, Woods, Conover, & Feix, 2003), there is relatively little information in the field about their diagnostic utility in student accommodation-seeking circumstances.

In spite of the limited knowledge base surrounding learning disability assessment, it does appear that external incentive, in the form of access to specialized resources as mentioned above, is sufficient to motivate students to feign or exaggerate mental health concerns or learning disabilities (Osmon, Plambeck, Klein, & Mano, 2006). As a result, the assessment of potential malingering by way of MDMs for this population is long overdue. The use of these measures may also be helpful in understanding students' performances on other tests, as learning disorder batteries traditionally include academic, personality and intelligence tests.

For example, evidence in other settings demonstrates that MDMs can substantially improve understanding of scores on impairment measures. Indeed, several studies have reported that malingering on some detection measures explained approximately five times more of the variance in composite neuropsychological test scores than traumatic brain injury severity (Green, Rohling, Lees-Haley, & Allen, 2001; Rohling & Demakis, 2010). This indicates that effort in

testing batteries may account for a large portion of a student's score, perhaps even more than a potentially present learning disability, and should be measured or accounted for in some way.

What is Malingering?

The DSM–IV–TR (American Psychiatric Association, 2000) defines malingering as the "intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs" (p. 739). Rogers et al. (1993) characterized malingering as manifesting in one of two ways: as an exaggeration of mental health disorder symptoms, and/or as a denial of lack of capability (lack of personal effort). In the field of neuropsychology, malingering assessment has been routine for many years and has resulted in the proposition of specific, commonly used criterion for diagnostic purposes (e.g., Slick, Sherman, & Iverson, 1999) which will be discussed below.

Malingering has traditionally been observed and measured in criminal/forensic/legal and neuropsychological settings where personal gain can be considerable. In general, forensic psychologists and neuropsychologists report that base rates for malingering among different populations varies widely, from 29% in personal injury cases (where financial incentives are powerful and always present) and 8% in medical cases to 19% in criminal cases (Mittenberg, Patton, Canyock, & Condit, 2002; Sullivan et al., 2007).

As there are potentially high rates of malingering in medical/personal injury and criminal practice, researchers in these two areas have contributed the most malingering detection measure development. This development, however, is based on malingering in different contexts and raises the question of how symptom distortion is present, as secondary gain incentives can be very different for these two populations (Green et al., 2001). For example, financial gain has

been noted as a primary malingering motivation, such as with personal injury claims and similar litigation (Larabee, 2003a), whereas avoidance of criminal responsibility and manipulation are primary malingering motivations in most forensic criminal populations (Hall, Poirier, & Thompson, 2008).

Malingering Detection Measures

In light of these two potential forms of malingering, detection has focused on assessment of effort, also known as effort testing (Osman & Mano, 2009) and unusual symptom report, also known as symptom validity testing (Nelson, Hoelzle, Sweet, Aribisi, & Demakis, 2010). Most MDMs were developed in legal/criminal and neuropsychological fields of study, and they both involve aspects of effort and symptom validity authentication. These measures have been well studied over the last 30 years and involve different forms of psychometric malingering measurement and data collection.

The four primary means of assessment include the following: (a) structured interviews, such as the *Structured Interview of Reported Symptoms* (SIRS; Rogers, Gillis, Bagby, & Dickens, 1991); (b) self-report symptom measures, such as the *Structured Inventory of Malingered Symptomology* (SIMS; Smith & Burger, 1997); (c) measuring testing effort via memory on a forced-choice task, such as with the *Test of Malingering and Malingering* (TOMM; Tombaugh, 1996), where malingering classification relies on underperformance or failure at levels below chance, as adopted from neuropsychological assessments such as the *California Verbal Learning Test – Second Edition* (Wolfe et al., 2010); and (d) through use of validity scales from forced-choice omnibus personality measures, such as the *F Scale* on the *Minnesota Multiphasic Inventory-2* (MMPI-2; Grieffenstein, Fox, & Lees-Haley, 2007).

Mittenberg et al. (2002), in explanation of a typical evaluation, noted that practicing neuropsychologists diagnosed malingering individuals using forced choice memory measures in 57% of cases and using validity scales from omnibus measures of personality in 38% of cases. Given the popular use of MDMs, researchers in the last decade have focused more extensively on measurement validity, allowing for the development of MDMs that have become more sophisticated and possess some moderate to excellent psychometric properties (Green, 2005; Rees, Tombaugh, Gansler, & Moczynski, 1998; Simon, 2007; Tombaugh, 1996).

Diagnostic considerations. Due to the multiple forms and methods of assessing malingering, Slick, Sherman, and Iverson (1999) provided what have become well-known and well-used criteria to assist in the diagnosis of malingered dysfunction, although some of their specific recommendations have collected criticism over the years (Boone, 2007). These criteria included three categories of malingering classification, labeled *possible*, *probable*, and *definitive*. Possible malingering is considered when external incentives are present and there is evidence from an individual's self-report that they did not put forth full effort or that they exaggerated symptoms. Probable malingering is considered when external incentives are present and there is evidence of two or more MDM failures that are non-forced choice. Definitive malingering is considered when external incentives are present and there is definite negative response, as evidenced by below-chance performance (failure) on one or more forced choice MDMs. Measures that lead directly to decisions (e.g., SIRS) are usually based on this structure, where clinicians use multiple data points that provide multiple scales with recommended cut points at which malingering is suspected, though not diagnosed.

Due to the very difficult nature of malingering detection, as respondents intentionally attempt to feign, distort, or deny symptoms, clinicians have been roundly cautioned to use

multiple measures or points of data, rather than one measure alone, to diagnose malingering (Rogers, 2008). This is evidenced in the Slick et al. (1999) criteria listed above, where a respondent must 'fail' at least two detection measures before probable malingering should be considered.

Symptom validity tests. As noted previously, one of two primary forms of malingering involves the intentional exaggeration or over-reporting of mental health symptoms or concerns. A key feature of symptom validation is comparing reported symptoms to expected diagnostic norms to look for inconsistencies and exaggeration (Rogers, 2008). For example, the F validity subscale on the MMPI-2 (F for Infrequency) has been an established standard indicating the validity of a respondent's profile (Graham, 2006). This scale is typically elevated when respondents endorse an improbably high number of symptoms, overly severe or bizarre symptoms for the nature of the condition they claim to have, symptoms that rarely occur together, and symptoms that rarely occur collectively as often as they endorse. These kinds of measures often rely on self-reports and usually contrast symptom severity, rarity, and bizarreness with previously established norms and other salient characteristics of the particular diagnostic category in question (Rogers et al., 1991).

Effort tests. The second form of malingering involves the denial of personal capability, where a respondent puts forth poor effort, resulting in a performance that is "less than" what would be observed in someone who was unimpaired. Poor effort can greatly affect test results, making a respondent appear more impaired or disordered, and a test administrator may not be able to detect levels of effort through clinical judgment alone. In order to gauge acceptable levels of effort, these MDMs require respondents to complete a task that relies on low face validity to mask its easiness. Based on normative performances of very impaired groups, these

tasks are not difficult, so if a respondent receives failing scores on an effort test, their performance is deemed questionable with regard to effort. Indeed, to malinger on effort measures, respondents are not forced to create complex symptom profiles and only need to get items wrong. These measures are commonly used in neuropsychiatric assessment as they are sensitive to neuropsychological impairment, and it is assumed that respondents without serious head injuries or cognitive impairment who malinger will display impairment much below expected levels, thereby raising suspicion of poor effort while testing (Merten, Bossink, & Schmand, 2007).

Memory tests. The most commonly assessed aspect of effort involves memory testing, where respondents are required to memorize and then recall very simple words or pictures. As with other effort tests, these memory tasks are made to appear more difficult than they are, providing malingerers an opportunity to feign impairment in general terms and without regard to a specific disorder. Further, the fact that the evaluator has chosen to administer a test of memory typically conveys to the malingerer that impaired persons must miss items; otherwise, the test would not be given.

The psychometrics of memory tests. Memory tests have become particularly salient in malingering detection due to their excellent psychometric properties. For example, as a group, memory tests display high specificity, where they are able correctly classify non-malingering respondents (Batt, Shores, & Chekaluk, 2008; Hubbard, 2008). They have been criticized, however, for the large variations in their sensitivity, or ability to properly classify malingerers, due to response bias (Gervais, Rohling, Green, & Ford, 2004). Response bias is the idea that when given psychological tests, respondents may perform according to how they believe they *should* perform, and not how they are truly capable (Gervais, Green, Allen, & Iverson, 2001),

which may interfere with malingering detection that attempts to distinguish true performance from altered performance.

Rates of sensitivity, or correct classification of malingering, across the majority of effort and memory tests is weak to moderate at best (Farkas, Rosenfeld, Robbins, & van Gorp, 2006; Greve, Binder, & Bianchin, 2009; Hubbard, 2008). As a result, it is even more important to understand and implement guidelines by Rogers (2008) and the Slick et al. (1999) criteria, both of which suggest the use of multiple points of data in order to diagnose malingering.

As a means of understanding the weak sensitivity of effort tests, recent researchers have theorized that effort and memory tests may have dissimilar underlying factor structures. In other words, these tests may tap into different facets of effort and memory and may indicate that respondents can malinger in many different ways. These factors potentially correlate with significantly different abilities, including declarative memory, attention, focus, and working memory (Merten et al., 2007). As the field continues to develop, if issues of response bias and dissimilar factor structures are any indication, these measures may require more detailed evaluation of their concordance, which in turn may result in more difficulty with consistent validation across populations.

Another important feature of memory tests is their potential for positive predictive value (PPV), or a test's ability to accurately predict actual malingerers based on test scores. A measure's PPV is difficult to determine in the absence of established base rates of malingering in a given population. In populations where base rates are known but very low, a measure's PPV may still be poor, due to the nature of a rare but present condition. This makes the need for base rate assessment even more salient when comparing and deciding which measures to use.

Four Malingering Detection Measures

Given the large number of MDMs available for use, the present study sought to examine and compare four tests that might provide the most efficient and accessible future testing battery for students at secondary education institutions. The first two MDMs are among some of the most widely used and studied measures of effort via memory tasks. They are designed specifically to gauge effort and are not typically part of a traditional learning disorder test battery, and thus they require more time and effort to administer on the part of a psychometrician. The second two MDMs are part of already-established tests of intelligence and personality that are very widely used in accommodation-seeking populations. These two MDMs are embedded in other measures but have a wide research base supporting their use as effort tests.

The Test of Memory and Malingering. The Test of Memory and Malingering (TOMM) is an effort measure that has been found to have near perfect specificity, in that it typically rules out all non-malingering respondents (Rees et al., 1998; Tombaugh, 1996). It appears to be insensitive to mild cognitive impairment, and those with mild mental retardation are able to produce scores above recommended cutoffs to produce "passing" rather than "failing" profiles (Simon, 2007). It has also been noted that it is insensitive to learning disabilities and ADHD in children and adults, where these groups produce passing scores (Tombaugh, 1996). Although the TOMM consists of three potentially administrable trials, researchers have determined that a "short" form (the initial TOMM trial) produces near similar specificity as all three trials combined, making it a short and efficient test to administer (Bauer et al., 2007; Gavett, O'Bryant, Fisher, & McCaffrey, 2005).

As with other MDMs, the TOMM has only modest sensitivity and struggles to accurately classify true malingerers all of the time (Gervais et al., 2004; Weinborn et al., 2003). The TOMM's excellent specificity and modest sensitivity may be explained by the extreme ease with which it is completed, where respondents are privy to the ease of the task and therefore made aware of nature of the test.

The Word Memory Test. The Word Memory Test (WMT) is another memory test that has gained notoriety, along with the TOMM, for its strong psychometric properties, ease of administration, and wide research base. In addition, it also has near perfect specificity (Gervais et al., 2004) and is also insensitive to learning disabilities and ADHD in children and adults (Green, 2005). The WMT has been widely validated in the detection of malingering in neuropsychological evaluations (Hartman, 2002), and was reported by Green (2005) to be highly successful at discriminating between poor effort/malingering and several types of genuine impairment. In several studies the WMT has been found to have better sensitivity than the TOMM when three WMT trials are compared to one TOMM trial, although researchers indicate that when three TOMM trials are used, sensitivity rates are roughly equal (Greiffenstein, Greve, Bianchini, & Baker, 2008).

The MMPI-2 F Scale. The Minnesota Multiphasic Inventory-2 (MMPI-2) is a measure of personality and general psychological functioning, and is among the most common psychological measures in the country (Graham, 2006). Elevated scores invalidate the measure and are typically indicative of symptom exaggeration across several domains. The F Scale score has been validated in many studies as having high specificity and moderate to high sensitivity in neuropsychological and forensic populations and is considered a very useful validity and

malingering screening tool (Heinly, 2005; Efendov, Sellbom, & Bagby, 2008; McCusker, Moran, Serfass, & Peterson, 2003).

WAIS Digit Span. The Digit Span (DS) is a total subscale score that has been age-corrected and gleaned from a respondent's performance on a memory subtest of the Wechsler Adult Scale of Intelligence (WAIS; Heinly, 2005). Even though DS has not traditionally been used as a malingering detection measure in learning disorder assessments, it has similar psychometric properties to other memory tests and has been increasingly studied and used as a measure of testing effort. It has been found to add incremental validity to malingering detection in litigious and neuropsychological assessment procedures (Axelrod, Fichtenberg, Millis, & Wertheimer, 2006; Etherton, Bianchini, Greve, & Heinly, 2005). Research indicates the DS also has high specificity and sensitivity when used in traumatic brain injury populations, even when moderate and severe TBI patients were included (Heinly, 2005).

The use of the DS has also been expanded, where Greiffenstein, Baker, and Gola (1994) created the Reliable Digit Span (RDS) test as separate means of gauging effort. Calculating the RDS involves summing the longest forward and backward digit strings of the DS subtest, where both trials were completed without error. Greiffenstein et al. (1994) found that the RDS had excellent specificity and moderate sensitivity in detecting poor effort, and other studies support these results (Mathias, Greve, Bianchini, Houston, & Crouch, 2002). As use of the RDS has become more popular and wide-spread, a meta-analytic review by Jasinski, Berry, Shandera and Clark (2011) examined 24 studies using either the RDS or DS (age-corrected scaled score) in malingering detection. They found no statistical differences between the RDS and DS in malingering classification and were both found to discriminate well between honest and responders and malingerers, with average weighted effect sizes of 1.34 and 1.08 respectively.

They both also demonstrated strong specificity and moderate sensitivity with cut rates at a score of six or below. As a result of their similarly strong psychometric properties, the DS was deemed a more ideal measure for the purpose of this study over the RDS, as it does not require additional calculations on the part of a test administrator.

Other uses of the DS in effort testing have also been explored, such as DS comparisons with WAIS Vocabulary scale scores, but this was found to have both poor specificity and sensitivity in use with learning-disordered populations (Harrison, Rosenblum, & Currie, 2010).

Malingering Detection Measure Concordance

As MDMs have been created, researched and applied in past years, many researchers have examined their utility or usefulness in traditional populations ranging from correctional and forensic institutions to neuropsychological and litigious settings (Edens, Poythress, & Watkins-Clay, 2007; Greiffenstein et al., 2008; Greve, Ord, Curtis, Bianchini, & Brennan, 2008; Sullivan & King, 2010). Indeed, the literature on the use of such measures in these settings is vast.

Malingering measures are typically compared in one of two ways, first across analogue samples, where participants are instructed to respond either "normally" or as "malingering" for some real or fictitious gain, known as instructed feigning. This allows researchers to determine if measures can identify and separate those participants who were told to fake symptomology or impairment versus those who were responding honestly. Since instructed malingering participants attempt to simultaneously exaggerate and hide that exaggeration, malingering measures rely on extreme reports (symptom validity) and assessment of effort to classify malingering.

Secondly, measures are also compared across convenience samples, where they are given to a set of actual clients in a particular setting without previous knowledge of which clients, if

any, are potential malingerers. This is known as natural feigning, and symptom and effort presentation may vary according to the setting and what kinds of secondary gains are present. Using multiple malingering measures across both of these samples allows researchers to compare how a measure detects natural versus instructed feigning.

Analogue and convenience samples also provide researchers with data on base rates of malingering in a population, as well as providing information about how a measure may classify malingering according to secondary gain. These different forms of comparison then determine measure failure concordance rates, or how well each measure is able to detect feigning. The use of several measures allows for the classification of malingering based on the failure of multiple tests. Hence, it is hypothesized that a malingering respondent will malinger consistently across tests, resulting in "failed" scores on at least two measures. Researchers then compare failure rates to determine if two or more tests are accurately identifying the same number of malingerers, because any one test of malingering may misclassify a respondent and result in the denial of services that are needed or required by law.

It should be noted that although multiple measure failures are helpful in classifying malingering, it is most helpful to have measures whose results are independent of one another (Rosenfeld, Sands, & van Gorp, 2000). In other words, if a respondent is falsely classified as malingering on Test A, that should not increase the likelihood of misclassification on Test B. The pragmatic usage of multiple tests, then, depends on the extent to which they misclassify the same individuals as malingering.

Present Study Hypotheses

This study attempted to address both the utility of MDMs and their concordance in a convenience sample of consecutively presenting students who were evaluated for learning

disabilities at a large, private, religiously-sponsored university in this population. The first analysis involved a comparison of four commonly used MDMs in terms of their classification rates. It was predicted that all four MDMs would independently classify malingering respondents, based on published cut scores and at expected base rates for an accommodation-seeking student population. The few studies that have been done cite a range of probable malingering base rates between 6-14%, although these are derived from various measures and depend on the particular characteristics of each university population studied (Harrison & Edwards, 2010; Harrison et al., 2007, Clayton, 2011). For the purposes of this study, the base rate was established at 10% due to recommendations from the above cited studies. Once calculated, the malingering classification rates were then used to create a malingering concordance analysis, or the degree to which each MDM agrees with other MDMs.

According to available research, the TOMM and WMT demonstrate malingering classification rates closest in approximation to the expected base rate of 10%. In the absence of a "Gold Standard" for symptom validity tests, both measures were used independently for the second analysis to calculate sensitivity, specificity, positive predictive value and negative predictive values for each of the other measures. It was proposed that each of the measures was likely to detect some malingering, although it was unknown if the characteristics of this specific population would change each measure's psychometric properties. These values were then compared to assess their inter-test correlations, with a null hypothesis stating that no significant differences would be found, based on the measures' similar psychometric properties in other studied populations.

A final examination of the data involved a qualitative comparison of the four MDMs to determine if they classified the same individuals as malingerers. Once again, research indicated

that these measures classify malingering at approximately the same rate in other populations, but given the lack of research with this specific population, it was simply unknown if these measures would classify the same individuals as honest responders or malingerers. As a result of the relative newness of this field, this analysis was largely exploratory in the absence of data on classification rates among malingering measures in university populations.

The principle hypotheses of this study are as follows. First, the TOMM, WMT, DS and F Scale measures would each yield suspected malingering classification rates similar to published base rates (between 8-14%). Second, based on a null hypothesis (due to a lack of previous research with this specific population), a malingering concordance analysis would reveal some degree of agreement among all MDMs. Third, specificity for each MDM would be high, but sensitivity would be moderate, and positive predictive value would be high with moderate negative predictive value, according to published research. Fourth, statistical comparison of the four measures would reveal no significant differences in their classification rates, based on a null hypothesis that, according to research, indicated similar agreement rates in other populations. Fifth, as a null hypothesis and due to an absence of previous research, a qualitative examination would reveal that no three measures would classify the same individuals as malingering.

Method

Procedure

The population in this study included students enrolled at least part-time at a large, private university and some students enrolled at a smaller business college associated with the university. Test administrators/evaluators were graduate student Ph.D. psychology interns or licensed psychologists, and testing sessions were scheduled for up to two hour periods. MDMs

were typically administered throughout the assessment process, and most students attended two or three testing sessions.

Participants

Archival data was gathered from a convenience sample of consecutively presenting students at a large private university ($N = 121$). Students who presented at the university's Accessibility Center were requesting evaluations for mental health concerns or learning disabilities (that interfered with their scholastic capabilities) in order to obtain accommodations during their college experience. These mental health issues spanned a wide spectrum, including depressive and anxiety disorders, ADHD, reading, math, and written expression disorders, and cognitive impairments. Archival data was gathered from a number of tests and included DS, F Scale, TOMM and WMT scores.

Measures

Although archival testing data on the MMPI-2 F scale and WAIS DS were available for several years prior to the implementation of the TOMM and WMT in the psychological test battery, only data from the first TOMM/WMT administration and forward were used in order to study the same sample characteristics across all tests.

The Test of Memory and Malingering. The Test of Memory and Malingering (TOMM) is an effort measure that asks a respondent to view 50 pictures of simply drawn common objects (animals, household items, etc.) that are presented a few seconds apart. After viewing all 50 pictures, a respondent is then shown a page with two pictures, one of which they viewed in the initial phase and one of which is a "distracter" picture; they are asked to point to the picture they were previously shown. After a correct response, the examiner provides positive feedback, "That is correct," and after an incorrect response, the examiner provides corrective

feedback, "No, it was this one" (pointing to correct picture). This is repeated for each of the 50 pictures. The process constitutes one trial, and there are two immediate trials given back to back, and one retention trial that may be given after 30 minutes that consists of only the second phase, where respondents are asked to identify the correct picture they previously viewed. Based on the speed of a respondent's answers, the first trial may take between 5-15 minutes to complete.

The TOMM consists of three separately administered trials, two immediate recall trials and one retention recall trial. As noted above, the first trial of the TOMM has been shown to produce equivalent outcomes when all three trials are compared, and it appears that using respondents' scores from the first trial will be sufficient to determine malingering classification rates. A recommended cutoff for malingering potential on the first trial of the TOMM, hereafter referred to as TOMM1, is any score under 45 (Tombaugh, 1996).

The Word Memory Test. The Word Memory Test (WMT) is administered via a computer-based program that presents respondents with simple word pairs (i.e. man – woman), at a third grade reading level, and asks them to memorize and recall these word pairs when shown after the initial presentation. In one recall task, they are given one half of the word pair and asked to identify the other. In another task, they are asked to recall as many of the word pairs as possible. Depending on the speed of the respondent, the administration takes anywhere from 10-20 minutes.

The WMT is similar to the TOMM, in that it focuses on the assessment of effort by testing memory. The WMT research base demonstrates that the first trial (immediate recall), hereafter referred to WMTIR, functions adequately as a screener for malingering, and thus should provide adequate malingering classification with a recommended cut score less than or equal to 82.5 (Green et al., 2002).

The MMPI-2 Infrequency Scale. The Minnesota Multiphasic Inventory-2 (MMPI-2) is a measure of personality and general psychological functioning, and is among the most common psychological measures in the country (Graham, 2006). The entire MMPI-2 is 567 questions long and must be administered in order to obtain the F-Scale scores. Given the length of the MMPI-2, it typically takes at least 45 minutes to complete.

The MMPI-2 contains a set of validity scales that includes the F-Scale, which stands for Infrequency. The scale includes dichotomous True/False questions about reported strange or unusual experiences, thoughts, sensations, paranoid ideations and antisocial attitudes and behaviors. Respondents who endorse many questions on this scale are reporting extremely high numbers of symptoms that are unusual or odd, and as a result, their higher than normal scores may indicate the presence of potential malingering.

The recommended cutoff for T-scores on the F scale has been variable according to the population studied, and although T scores above 70 are considered suspect for potential exaggeration, when detecting malingering in particular, a higher cutoff T score is recommended, typically at greater than or equal to 95 (Graham, 2006; McCusker, Moran, Serfass, & Peterson, 2003).

The WAIS Digit Span. The Wechsler Adult Intelligence Scale (WAIS) – Digit Span (DS) is a memory subtest on a measure of intellectual functioning. It assesses memory by asking respondents to repeat strings of numbers until they make consecutive mistakes asks respondents to listen to a series of numbers and repeat those numbers in the correct sequence requested. During the year these administrations took place, the testing battery was updated from the use of the WAIS – 3rd Edition (WAIS-III) to the WAIS – 4th Edition (WAIS-IV). Of the available 76 DS profiles, the WAIS-III was given to the first 50 participants and the WAIS-IV was used for

the last 26 participants. The WAIS-III DS asks respondents to repeat numbers both forward and backward, while the WAIS-IV asks respondents to repeat numbers forward, backward, and by ascending order from 1-10. Although these two tasks create different working memory demands, recent research indicated that the new WAIS-IV DS has parallel psychometric properties to the WAIS-III DS and matches its ability to detect malingering, even when compared to other MDMs, such as the WMT (Young, Sawyer, Roper, & Baughman, 2012).

During the administration of the DS, respondents complete one trial at a time, where a trial includes two sets of number sequences of the same length. They continue to repeat number sets that increase by one digit per trial until they fail one total trial or successfully reach the highest trial. The DS trials are finished after a respondents' failure, so length of administration depends on how many successive trials are passed. Given its lack of visual capacity or verbal knowledge requirements, outside of familiarity with the numbers 0-9, it is considered a good measure of memory and effort. Recommended cut scores for malingering detection are typically scores of less than 6.8 (Heinly, 2005; Jasinski et al., 2011).

Results

Comparison of Measure Classification Rates

There were different sample sizes for each of the four MDMs studied, due to missing data points arising from several factors. First, different test protocols were administered according to presenting concerns for each student, making some tests necessary and others not. Second, some students began certain tests as part of their psychological assessment but did not complete them due to interruptions in their schedules. If a valid score was available for any of the four measures, it was included in the analyses. This resulted in the sample sizes and population characteristics listed in Table 1.

In an analysis of gender differences using two-tailed t-tests, gender was not significant for the TOMM1, WMTIR, DS or total sample size, but did meet criteria for statistical significance on the F Scale. A more thorough examination revealed that even in the presence of extremely similar subsample sizes (49 male, 46 female), male F Scale scores had *M* = 56.81, *SD* = 13.45, and female F Scale scores had Qualitative Comparisons = 63.91, *SD* = 19.59. The initial observed difference between the F Scale and the other tests is that the three MDMs are measures of effort via memory (lack of capability), while this measure is an embedded validity scale that assesses over-reporting of symptoms (exaggeration). It is unknown if these different formats for assessing malingering are affected by gender bias, social gender roles and norms, or other events or expectations.

Table 1. *Demographic characteristics*

	TOMM1	**WMTIR**	**F Scale**	**DS**	**Total**
N	117	105	95	76	121
Female gender *n* (% sample)	53 (45.3)	48 (45.7)	46 (48.4)	37 (48.7)	55 (45.5)
Male gender *n* (% sample)	64 (54.7)	57 (54.3)	49 (51.6)	39 (51.3)	66 (54.5)
t-test of gender	.29	.38	-2.05*	1.68	-.11
Age (years) *M (SD)*	26.3 (7.7)	26.3 (7.6)	26.6 (7.9)	25.6 (6.5)	26.2 (7.5)

*$p<.05$

In order to receive a "pass" on each measure, a respondent's score needed to indicate the presence of honest effort and honest responding. Hence, based on the cut-offs referenced above, malingering was operationalized as a score that fell below 45 for the TOMM1, below or equal to 82.5 for the WMTIR, at or above 95 for the F Scale, and at or below 6.8 for the DS. Once scores were classified as either pass or fail, each malingering measure's failure rate was calculated. The

predicted rate of correct malingering classification for each measure was also calculated based on the sample size, or number of administered protocols, at a predicted base rate of 10%. A summary of the classification rates are presented in Table 2.

Table 2. *Classification Rates of Malingering*

Test	*N*	Fails	Malingering Classification Rate	Predicted Fails at 10% BR	Chi Square Goodness of Fit Test
TOMM1	117	13	11.11%	12	0.88
WMTIR	105	10	9.52%	11	0.86
F Scale (95)	95	4	4.21%	10	0.06
F Scale (80)	95	13	13.68%	10	0.40
DS	76	13	17.11%	8	0.09

According to the first proposed hypothesis, the TOMM1, WMTIR, F Scale and DS should have displayed roughly the same ability to identify malingering respondents. According to the number of identified fails, there was a moderate, though not statistically significant, range of malingering classification rates. The DS resulted in the greatest amount of fails, 13 available protocols out of an expected eight (17.11%), which is considerably higher than any of the other classification rates and outside the expected range. The TOMM1 and WMTIR classification rates fell at 11.11% and 9.52%, respectively, which are very close to the expected 10% base rate of malingering.

The F Scale ($\geq$ 95 cutoff score) identified the least amount of fails (4.21%), which is a failure rate considerably below the expected base rate and outside the 8 – 14% potential malingering range found in the research. Due to this unexpectedly low percentage, the classification rate for a different cut score was calculated based on recommendations that scores between 70 and 90 should be examined for potential exaggeration (Graham, 2006). A median

cut score of 80 (> 79) was used after a search of the literature specified that T scores as low as 65 have been used to detect malingering with excellent specificity and sensitivity, although most studies recommend cut scores above 75 (Grieffenstein et al., 2007). The new F Scale (≥ 80 cutoff score) was found to substantially increase the failure detection rate at 13.68%, which falls inside the 8-14% expected range.

Table 3. *Agreement rates between TOMM1 and [WMTIR, F Scale 95, F Scale 80, and DS]*

TOMM1	WMTIR (*n*=101)		F Scale 95 (*n*=92)		F Scale 80 (*n*=92)		DS (*n*=74)	
	Fail	Pass	Fail	Pass	Fail	Pass	Fail	Pass
Fail (%)	3 (3)	10 (9.9)	0 (0)	10 (10.9)	3 (3.3)	7 (7.6)	1 (1.4)	4 (5.4)
Pass (%)	7 (6.9)	81 (80.2)	3 (3.3)	79 (85.9)	9 (9.8)	73 (79.3)	12 (16.2)	57 (77.0)

The concordance analyses were calculated once malingering classification rates were completed, and each measure was compared to every other measure, including both cut score profiles of the F Scale. For ease of readability, the measures are presented in groupings, first by comparisons of the TOMM1 with all tests (Table 3), second by comparisons of the WMT and all tests minus the TOMM1 (Table 4), and third by comparisons of the remaining tests, DS/F Scale 95 and DS/F Scale 80 (Table 5).

The agreement rate between the TOMM1 and WMTIR was 83.2% (summing the joint pass/pass and fail/fail cells in Table 3), and they classified the same three respondents as malingering. Similar results were found for the TOMM1 and F Scale 80 (82.6%). The TOMM1 and DS had a lower agreement rate of 78.4%, which was surprising, given that the DS had a high malingering classification rate of 17.11% and the TOMM1 a rate of 11.11%. The agreement rate between the TOMM1 and F Scale 95 was highest at 85.9%. They did not classify any of the

same respondents as malingering, and although they had a high agreement rate based on joint pass/pass, this is likely due to the fact that the F Scale 95 has the lowest base rate of all the four measures.

Table 4. *Agreement rates between WMTIR and [F Scale 95, F Scale 80, and DS]*

WMTIR	F Scale 95 (*n*=92)		F Scale 80 (*n*=92)		DS (*n*=74)	
	Fail	Pass	Fail	Pass	Fail	Pass
Fail (%)	1 (1.2)	8 (9.3)	3 (3.5)	6 (7.0)	2 (2.9)	3 (4.4)
Pass (%)	3 (3.5)	74 (86.0)	9 (10.5)	68 (79.1)	10 (14.7)	53 (77.9)

Comparisons of the WMTIR and F Scale 95 resulted in an agreement rate of 87.2% (Table 4). This is slightly higher than the WMTIR and F Scale 80 agreement rate of 82.6% but may be due to low F Scale 95 malingering classification rate of 4.21%, where there is a smaller chance for false positive misclassification. The WMTIR and DS comparison rate was 80.8%, which is the lowest WMTIR comparison.

Table 5. *Agreement rates between [F Scale 95 and DS] and [F Scale 80 and DS]*

F Scale 95	DS (*n*=74)		F Scale 80	DS (*n*=74)	
	Fail	Pass		Fail	Pass
Fail (%)	1 (1.7)	1 (1.7)	Fail (%)	1 (1.7)	5 (8.3)
Pass (%)	9 (15.0)	49 (81.7)	Pass (%)	9 (15.0)	45 (75.0)

The agreement rate between the F Scale 95 and the DS was 83.4%, and the rate between the F Scale 80 and the DS was 76.7% (Table 5), the lowest among all the measure comparisons.

The cut scores resulted in significantly different agreement rates, where the F Scale 95 displayed more agreement with all three MDMs than did the F Scale 80 (Table 6).

Table 6. *Cut Score differences between the F Scale 95 and F Scale 80*

	F Scale 95	F Scale 80
TOMM1	85.90%	82.60%
WMTIR	87.20%	82.60%
DS	83.40%	76.70%

Psychometric Properties and Inter-test Correlations

In the absence of reliable diagnostic accuracy for malingering in this student population, research suggests that the TOMM and WMT offer the most accurate classification rates among MDMs and can be used in psychometric calculations (Mossman, Wygant, & Gervais, 2012; O'Bryant & Lucas, 2006). As a result, each measure was used independently to calculate the sensitivity, specificity, positive predictive and negative predictive values of the other four MDMs in the study (Table 7, Table 8).

Table 7. *Sensitivity, Specificity, Positive Predictive Value and Negative Predictive Value using the TOMM1*

	WMTIR	F Scale 95	F Scale 80	DS
Sensitivity	0.23	0.00	0.30	0.20
Specificity	0.92	0.96	0.89	0.83
PPV	0.30	0.00	0.25	0.08
NPV	0.89	0.89	0.91	0.93

When the TOMM1 was used as the "Gold Standard" of malingering classification, the specificity was excellent (.83 - .92), where each measure correctly classified honest responders

most of the time (Table 7). Correspondingly, they also displayed excellent NPV (.93 - .89), or a high likelihood that any passing score (classified as probably non-malingering) would be an actual honest responder. However, the sensitivity for each measure was poor, falling at .30 and below, where they failed to correctly identify potentially malingering respondents in the sample. The positive predictive power for each measure was also poor, again falling at .30 and below, indicating the poor likelihood that any predicted malingering respondent would be an actual malingerer. In fact, the F Scale 95 displayed no sensitivity or PPV at all, and the PPV of the DS measure fell near zero (.08). Although each measure had poor sensitivity and PPV when the TOMM1 was used, there was not a wide spread among each measure's psychometric properties; or in other words, each measure performed relatively the same with this population.

Table 8. *Sensitivity, Specificity, Positive Predictive Value and Negative Predictive Value using the WMTIR*

	TOMM1	F Scale 95	F Scale 80	DS
Sensitivity	0.30	0.11	0.33	0.40
Specificity	0.89	0.96	0.88	0.84
PPV	0.23	0.25	0.25	0.17
NPV	0.92	0.90	0.92	0.95

When the WMTIR was used as the "Gold Standard" of malingering classification (Table 8), the specificity again was excellent (.84 - .96), and correspondingly, the NPV was excellent (.90 - .95). The sensitivity for each measure was also poor (.11 - .40), although rates were slightly higher under the WMTIR than when the TOMM1 was used. Positive predictive values were poor, ranging from .17 to .25, but when the WMTIR was used, the PPV range was truncated. Once again, although each measure had poor sensitivity and PPV, there was not a wide spread among each measure's psychometric properties.

The two embedded measures (F Scale 80/F Scale95 and DS) were also compared with the two measures that are dedicated to the detection of malingering (TOMM1 and WMTIR) using their respective positive predictive values. The PPV, or probability that a test will accurately identify malingering, is considered one of the most important aspects of a MDM, although it must be balanced in this population with acceptable negative predictive values, as university disability service providers want to accurately identify dishonest responders but dislike denying accommodations to students in need. With that in mind, overall positive predictive values were highest when the WMTIR was used as the "Gold Standard." The F Scale 95 had a higher PPV when used with the WMTIR (.25) as opposed to the absence of PPV when used with the TOMM1 (0.0). The F Scale 80 had identical PPV with both the TOMM1 and WMTIR (.25).

The DS displayed the lowest PPVs (TOMM1 = .08, WMTIR = .17) and the highest malingering classification rate (17.11%, Table 2) out of all the MDMs. The DS cutoff score (DS <7) was used in calculations due to recommendations from a strong research base that indicated in neuropsychological and other forensic populations, the DS is an excellent measure of effort. Its low PPV with this population, however, suggests that it may not provide the most accurate malingering assessment, especially when compared to a dedicated MDM.

The F Scale 95 also displayed weaker psychometric properties than the F Scale 80 (0.0 versus .25) when the TOMM1 was used as the "Gold Standard." Due to the lower PPV of the DS, if the F Scale 95 is used as a malingering measure, it appears that the addition of a dedicated MDM (TOMM or WMT) would increase the probability of more accurate classification over the use of the DS.

As a means of assessing the degree to which these four measures may have been measuring similar constructs, or may have provided overlapping information, Pearson

correlations were calculated between each of the four tests using corresponding raw scores for each measure (Table 9). The highest and indeed only significant inter-test correlation (.46) was between the WMTIR and TOMM1, suggesting that these two measures may be providing similar information about malingering rates in this population.

Table 9. *Intercorrelations among Malingering Detection Measures*

	TOMM1	WMTIR	F Scale 95	F Scale 80	DS
TOMM1	-	(102)	(92)	(92)	(76)
WMTIR	0.46**	-	(86)	(86)	(68)
F Scale 95	-0.12	-0.13	-	-	(60)
F Scale 80	-0.12	-0.13	-	-	(60)
DS	0.02	0.18	-0.04	-0.04	-

Values in parentheses are cell sizes.
**$p < .01$

Qualitative Comparisons

In previous portions of this study, the TOMM1, WMTIR, F Scale 95, F Scale 80 and DS malingering classification rates were calculated and compared to see if any measures displayed an incremental advantage at detecting failed protocols. Their failure rates were also compared to see if each measure was capturing the same information or the same type of malingering.

In the third analysis of this study, these four measures were qualitatively compared to determine if they classified the *same respondents* as malingering. The cross-tabulation matrices listing agreement rates among measures (Tables 3, 4, and 5) reveal how each MDM compares to the others in terms of classifying the same respondents. For example, the TOMM1 failed three of same respondents as the WMTIR, but did not capture any similar failed respondents with the DS. As noted above, Rogers (2008) and Slick et al. (1999) both recommend using multiple test

fails in order to accurately classify malingering. If there are truly different kinds of malingering (lack of capability/low effort versus over-reporting/exaggeration of symptoms), however, it is important to have measures whose results are capture different facets of the malingering construct, and are therefore somewhat independent of one another (Rosenfeld, Sands, & van Gorp, 2000).

In order to determine if and how *all four measures* classified the same respondents, the only respondents included were those whose data points were available for all four tests, resulting in $n = 54$. When all four data points for each respondent were examined side by side, it was revealed that no more than two measures identified a single respondent as malingering, even using the higher classification rate of the F Scale 80. In other words, no three measures in this sample identified the same respondent as malingering.

Given these results, another qualitative data examination was performed, where rather than deleting a profile with missing data points, all respondent profiles were compared side by side, with missing data points blacked out. It was then discovered that two respondents out of the total population size ($N = 121$) received failing scores on three measures: the TOMM1, WMTIR, and F Scale 80 (the DS score was not even available, as the WAIS was not administered in these two cases).

Discussion

The first part of this study sought to calculate malingering classification rates for each of the four MDMs (TOMM1, WMTIR, DS and F Scale). It was hypothesized, in the absence of a strong research base using these measures with this population, that each MDM would classify malingering at the suggested base rate of 10%. The TOMM1 and WMTIR produced classification rates of 11.11% and 9.52%, respectively, which are very close to the predicted base

rate. The DS classification rate was somewhat higher than predicted, even though it was based on solid, well-researched cutoff scores.

The F Scale classification rate (at a cutoff score of 95) was significantly lower than predicted, and as a result, another cutoff score of 80 was established and resulted in a classification rate that better approximated the predicted base rate. The necessity for cutoff score changes may be due to the characteristics of this population, but also to the fact that while use of the MMPI-2 F scale in malingering research is very common, cutoff scores span a very wide range (i.e. scores above 69 may indicate "exaggeration," while scores above 90 may be "faking bad") and there is no gold standard for F Scale malingering interpretation. It is also true that given the demand characteristics of this specific population (ability to function at a university scholastic level), high F Scale scores based on true psychiatric impairment is much less likely than other populations where malingering may be present. As a result, lowering cut scores to increase sensitivity on the F Scale may be more feasible, as the risk of mislabeling true psychopathology as malingering is reduced.

The second hypothesis of this study suggested that there would be some agreement of malingering detection rates among MDMs, as a null hypothesis, in a malingering concordance analysis (a variant of the differential prevalence design that compares classification rates in the absence of definite external criteria for malingering). The lowest agreement rate was moderately high, at 76.7% between the DS and F Scale 80, and the highest agreement rate, between the TOMM1 and F Scale 95, was 85.9%. It was interesting to note that the TOMM1 and F Scale 95 did not classify any of the same respondents as malingering; therefore their high agreement rate was based solely on mutual passing classifications. This may have resulted from the fact that the F Scale 95 only classified four respondents as failing and was therefore a more conservative

measure of malingering than the TOMM1, which classified 13 respondents as failing. In spite of high malingering classification rates among these MDMs, it is important to note that there are inherent problems with low base rates of malingering in this population (O'Bryant & Lucas, 2006; Mossman, Wygant, & Gervais, 2012). Indeed, these measures by virtue may have high classification rates based on true responders while failing to capture true malingerers, as is demonstrated by the excellent NPVs and poor PPVs in this study.

The third hypothesis proposed that specificity for each MDM would be high, but sensitivity would be moderate. Specificity for each measure was indeed high when both the TOMM1 and WMTIR were used as comparisons, with rates above .83 for each measure. Sensitivity was also lower than expected for each measure, and even fell below moderate levels, with the highest at .40 indicating poor sensitivity all around. When used in other populations, these measures have demonstrated only moderate sensitivity as well, and their underperformance in this population may indicate that each measure is assessing aspects of malingering, but also concomitant characteristics of this population.

It was also hypothesized that PPV and NPV would be moderate, according to published research. PPV was in fact rather poor, with rates falling at and below .30. PPV for the F Scale 95 was completely absent when the TOMM1 was used as the "Gold Standard," though it increased somewhat to .25 with the WMTIR. NPVs were high for each measure, falling above .89. Although malingering measure usage relies heavily on PPV, in situations such as this where base rates are unknown, unsure or varied, strong specificity is of utmost importance in order to reduce false positive errors.

The fourth hypothesis of the study suggested that a statistical comparison of the four measures would reveal no significant differences in their ability to detect malingering, based on a

null hypothesis that, according to research, indicated similar agreement rates in other populations. Intercorrelation statistics revealed that only the TOMM1 and WMTIR were significantly correlated (.46), suggesting that these two measures may be providing similar information about malingering rates in this population. This may be an important factor to consider when creating or implementing a testing battery, especially since these two measures have similar psychometric properties within this population.

The last hypothesis involved a qualitative examination of respondents who were potentially identified as malingering by more than two (and up to four) tests. The null hypothesis was that no individual respondent would be classified as malingering by more than two measures. When all four data points were present (n=54), the null hypothesis held true, as no respondents were classified malingering by more than two tests.

When all data points were examined without exclusion, it was found that two respondents were classified as malingering by three of the tests (DS was the missing data point). This may be due to several factors that influenced the progression of the testing battery. For example, the TOMM and WMT were not administered at any specific point in the evaluation process, but if they were administered first and resulted in one or more failed scores, the evaluator may have declined to give any more tests in the face of potential malingering. In another example, if the TOMM and WMT were administered first and resulted in failed scores, the evaluator may have decided to give an MMPI-2 to rule out personality or mental health disorders, or may have declined to give a WAIS in the face of potential malingering (requiring less evaluator rigor, time and money).

It appears that each of these four measures do share the ability to detect malingering in its different forms (lack of capability/low effort and over-reporting/exaggeration of symptoms).

There also appeared to be moderately high agreement among the MDM's classification rates, but each measure also displayed enough independence that they appeared to provide additive information, with the exception of the TOMM1 and WMTIR scales. These two measures were significantly correlated and may provide overlapping information in test batteries. Three of the measures displayed excellent psychometric properties across the board (TOMM1, WMTIR, and F Scale 95), while the DS and F Scale 80 presented with only moderate PPV. When the qualitative comparison was made, however, the F Scale 80 classified two respondents as malingering in tandem with the TOMM1 and WMTIR, providing incremental validity for those particular protocols.

In any case, the F Scale appeared to provide robust PPV at a cut score of 95 and additive validity at a cut score of 80, and the DS appeared to provide strong NPV and excellent sensitivity and specificity. As these two measures are embedded in traditional learning disorder testing batteries, and since the TOMM and WMT have near-identical psychometric properties and provide overlapping information, it may be pragmatic to employ the use of only one of these measures in order to assist in the evaluation of suspected malingering with accommodation-seeking students, although more research in this field is definitely needed.

Limitations

There are several relevant limitations to this study that are typical in early research field development. First, base rate approximation is an issue, due to the fact that malingering in this population has only recently been studied and may be relatively low. This affects each of the four measures' ability to adequately detect malingering populations. In addition, it has been demonstrated in past studies that college students are more sensitive to sophisticated (extremely low face validity) versus unsophisticated (high face validity) methods of feigning, so by virtue of

this particular quirk of the research population, these measures may not adequately detect malingering across this sample.

A second limitation involves concerns relative to the incentive in an accommodation-seeking population to malinger. Some researchers have argued that students who are evaluated for learning diagnoses are more prone to avoid failure, rather than seek specific accommodations, and this may incentivize them to only mildly exaggerate their symptom profiles, resulting in suboptimal effort that may nevertheless result in sub-threshold cut scores. A third limitation involves the process of gathering data from a convenience sample. Selection characteristics or issues of random heterogeneity in the sample are important, as particular groups of people may be more prone to service-seeking than others. Also, some students may present with disorder-specific questions that alter or distort their effort presentation, relative to mathematics disabilities, ADHD, or mood disorder concerns, for example.

A fourth limitation is the problem of response bias or experimenter expectancies. Some researchers have proposed the idea that when administering malingering detection measures, experimenters influence how the participant responds in unknown or unintentional ways. This is particularly important on the TOMM and DS, as the examiner provides immediate feedback on correct versus incorrect answers. This personal feedback interaction may significantly alter a respondent's ability to successfully malinger on these tests. A final limitation involves external validity in general. Research and field development in the area of university student malingering potential is minimal, and although external application of these results is premature, it will be important to address the generalizability of malingering and its detection in student populations in future research.

Summary and Future Directions

As the need for malingering detection in university populations increases, the evaluation of appropriate MDMs and of their usefulness can be very useful in determining the pragmatic choices of which measures to use according to personalized financial, staffing, and time constraints. According to this study, the most practical suggestion would involve the use of the embedded measures (DS, F Scale) to assist in the screening of malingering, and the employment of one of the two dedicated MDMs (TOMM, WMT) may be useful in meeting the Slick et al. (1999) criteria for diagnosing malingering if warranted.

For the F Scale in particular, the two different cut scores are recommended for two different situations. First, with the use of other MDMs a cut score of 80 may provide the most helpful information to rule out false positives while indicating the presence of possible malingering. Second, in the absence of other data points, if malingering is suspected, a cut score of 95 is recommended as a more guarded and cautious estimation of alleged malingering. This is due to the potentially biased nature of labeling someone as malingering and the possibility that someone who may genuinely need services does not get them, based only on one criterion.

In summary, the field of malingering learning disorders in accommodation-seeking university populations is fairly new, and more research is needed to determine base rates across a wide range of universities and colleges. In particular, there is a great need for measures that are specific for malingering detection within the LD population, as the measures currently in use were developed in a neuropsychiatric environment and do not demonstrate strong, comprehensive psychometric properties with this specific population.

References

Advokat, C. D., Guidry, D., & Martino, L. (2008). Licit and illicit use of medications for Attention-Deficit/Hyperactivity Disorder in undergraduate college students. *Journal of American College Health, 56(6),* 601-606.

American Psychiatric Association (APA). (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. Washington, DC

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

Batt, K., Shores, E.A., & Chekaluk, E. (2008). The effect of distraction on the Word Memory Test and Test of Memory and Malingering performance in patients with a severe brain injury. *Journal of the International Neuropsychological Society, 14(6)*, 1074-1080.

Bauer, L., O'Bryant, S. E., Lynch, J. K., McCaffrey, R. J., & Fisher, J. M. (2007). Examining the Test of Memory Malingering trial 1 and Word Memory Test Immediate Recognition as screening tools for insufficient effort. *Assessment, 14(3)*, 215-222.

Bianchini, K. J., Etherton, J. L., Greve, K. W., Heinly, M. T., & Meyers, J. E. (2008). Classification accuracy of MMPI-2 validity scales in the detection of pain-related malingering: A known-groups study. *Assessment, 15(4)*, 435-449.

Blanchard, D. D., McGrath, R. E., Pogge, D. L., & Khadivi, A. (2003). A comparison of the PAI and MMPI-2 as predictors of faking bad in college students. *Journal of Personality Assessment, 80(2),* 197-205.

Boone, K. (2007). A reconsideration of the Slick et al. (1999) criteria for malingered neurocognitive dysfunction. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment* (pp. 29-49). New York, NY: The Guilford Press.

Clayton, S. (2011). Malingering detection among accommodation-seeking university students. *Dissertation Abstracts International: Section B: The Sciences and Engineering, Vol 71*(12-B), 2011. pp. 7719.

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.

Crank, J. N., & Deshler, D. D. (2001). Disability eligibility issues and university student assessment outcomes. *Journal of Vocational Rehabilitation, 16*, 217-226.

Edens, J. F., Poythress, N. G., & Watkins-Clay, M. M. (2007). Detection of malingering in psychiatric unit and general population prison inmates: A comparison of the PAI, SIMS, and SIRS. *Journal of Personality Assessment,88(1)*, 33-42.

Efendov, A. A., Sellborm, M., & Bagby, R. M. (2008). The utility and comparative incremental validity of the MMPI-2 and Trauma symptom Inventory validity scales in the detection of feigned PTSD. *Psychological Assessment, 20(4)*, 317-326.

Etherton, J. L., Bianchini, K. J., Greve, K. W., & Heinly, M. T. (2005). Sensitivity and specificity of Reliable Digit Span in malingered pain-related disability. *Assessment, 12(2)*, 130-136.

Farkas, M.R., Rosenfeld, R., Robbins, R., & van Gorp, W. (2006). Do tests of malingering concur? Concordance among malingering measures. *Behavioral Sciences and the Law, 24(5),* 659-671.

Frazier, T. W., Frazier, A. R., Busch, R. M., Kerwood, M. A., & Demaree, H. A. (2008). Detection of simulated ADHD and Reading Disorder using symptom validity measures. *Archives of Clinical Neuropsychology, 23*, 501-509.

Gavett, B. E., O'Bryant, S. E., Fisher, J. M., & McCaffrey, R. J. (2005). Hit rates of adequate performance based on the Test of Memory Malingering (TOMM) Trial 1. *Applied Neuropsychology, 12(1)*, 1-4.

Gervais, R. O., Ben-Porath, Y. S., Wygant, D. B., & Green, P. (2007). Development and Validation of a Response Bias Scale (RBS) for the MMPI-2. *Assessment, 14*, 196-208.

Gervais, R.O., Green, P., Allen III, L. M., & Iverson, G. L. (2001). Effects of coaching on symptom validity testing in chronic pain patients presenting for disability assessments. *Journal of Forensic Neuropsychology*, 2(2), 1-19.

Gervais, R. O., Rohling, M. L., Green, P., & Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Archives of Clinical Neuropsychology, 19*, 475-487.

Graham, J. R. (2006). *MMPI-2: Assessing Personality and Psychopathology* (Fourth ed.). New York: Oxford University Press.

Green, P. (2005). *Green's Word Memory Test*. Edmonton: Green's Publishing Inc.

Green, P., Lees-Haley, P. R., & Allen, L. M. (2002). The Word Memory Test and the validity of neuropsychological test scores. *Journal of Neuropsychology, 2*(3), 97-124.

Green, P., Rohling, M. L., Lees-Haley, P. R., & Allen, L. M., III. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury, 15(12*), 1045-1060.

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

Greiffenstein, M. F., Fox, D. D., & Lees-Haley, P. R. (2007). The MMPI-2 in detection of non-credible brain injury claims. In K. B. Boone (Ed.)*, Assessment of Feigned Cognitive Impairment: A Neuropsychological Perspective* (pp. 210-235). New York: Guilford Press.

Greiffenstein, M. F., Greve, K. W., Bianchini, K. J., & Baker, W. J. (2008). Test of Memory and Malingering and Word Memory Test: A new comparison of failure concordance rates. *Archives of Clinical Neuropsychology, 23(7-8)*, 801-807.

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

Greve, K. W., Ord, J., Curtis, K. L., Bianchini, K. J., & Brennan, A. (2008).Detecting malingering in traumatic brain injury and chronic pain: A comparison of three forced-choice symptom validity tests. *The Clinical Neuropsychologist, 22(5)*, 896-918.

Hall, H. V., Poirier, J. G., & Thompson, J. (2008). Detecting malingering and deception in forensic evaluations. In H. V. Hall (Ed.), *Forensic psychology and neuropsychology for criminal and civil cases* (pp. 93-130). Boca Raton: CRC Press.

Harrison, A. G. (2006). Adults faking ADHD: You must be kidding! *ADHD Report*, *14*, 1–7.

Harrison, A. G., Edwards, M. J., & Parker, K. (2007). Identifying students faking ADHD: Preliminary findings and strategies for detection. *Archives of Clinical Neuropsychology, 22*, 577-588.

Harrison, A. G., & Edwards, M. J. (2010). Symptom exaggeration in post-secondary students: Preliminary base rates in a Canadian sample. *Applied Neuropsychology, 17(2)*, 135-143.

Harrison, A.G., Rosenblum, Y., & Currie, S. (2010). Examining unusual digit span performance in a population of postsecondary students assessed for academic difficulties. *Assessment, 17*(3), 283-293.

Hartman, D. E. (2002). The unexamined lie is a lie worth fibbing: Neuropsychological malingering and the Word Memory Test. *Archives of Clinical Neuropsychology, 17,* 709-714.

Heinly, M. T. (2005). WAIS Digit Span-based indicators of malingered neurocognitive dysfunction: Classification accuracy in traumatic brain injury. *Assessment, 12(4)*, 429-444.

http://ada.gov (2008). Americans with Disabilities Act of 1990, As Amended [Government publication of amendments made to the 1990 ADA]. Retrieved from http://www.ada.gov/pubs/ada.htm

Hubbard, K.L. (2008). Feigning cognitive deficits among psychiatric inpatients: Validation of three measures of cognitive malingering. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 68(10-B)*, 6966.

Jachimowicz, G., & Geiselman, R. E. (2004). Comparison of ease of falsification of attention deficit hyperactivity disorder diagnosis using standard behavioral rating scales. *Cognitive Science Online*, *2*, 6–29.

Jasinski, L. J., Berry, D. T. R., Shandera, A. L., & Clark, J. A. (2011). Use of the Wechsler Adult Intelligence Scale Digit Span subtest for malingering detection: A meta-analytic review. *Journal of Clinical and Experimental Neuropsychology, 33(3)*, 300-314.

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

Larrabee, G. J. (2003b). Detection of symptom exaggeration with the MMPI-2 in litigants with malingered neurocognitive dysfunction. *The Clinical Neuropsychologist, 17(1),* 54-68.

Latham, P. H. (2005). Learning disabilities and the law: After high school: An overview for students. Retrieved December 15, 2009 from http://www.ldanatl.org/aboutld/adults/civil_rights/law.asp

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

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

McCabe, S. E., Teter, C. J., & Boyd, C. J. (2006). Medical use, illicit use, and diversion of prescription stimulant medication. *Journal of Psychoactive Drugs, 38(1),* 43-56.

McCann, B. S., & Roy-Byrne, P. (2004). Screening and diagnostic utility of self-report attention deficit hyperactivity disorder scales in adults. *Comprehensive Psychiatry*, *45*, 175–183.

McCusker, P. J., Moran, M. J., Serfass, L. & Peterson, K. H. (2003). Comparability of the MMPI-2 F(p) and F Scales and the SIRS in clinical use with suspected malingerers. *International Journal of Offender Therapy and Comparative Criminology, 47(5)*, 585-596.

McGuire, J. M. (1998). Educational accommodations: A university administrator's view. In M.Gordon & S.Keiser (Eds.) , *Accommodations in higher education under the Americans With Disabilities Act (ADA): A no-nonsense guide for clinicians, educators, administrators, and lawyers* (pp. 20– 45). DeWitt, NY: GSI Publications.

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

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

Mossman, D., Wygant, D. B., & Gervais, R. O. (2012, April 30). Estimating the accuracy of neurocognitive effort measures in the absence of a "Gold Standard". *Psychological Assessment*. Advance online publication. doi: 10.1037/a0028195

Nelson, N. W., Hoelzle, J. B., Sweet, J. J., Arbisi, P. A., & Demakis, G. J. (2010). Updated meta-analysis of the MMPI-2 Symptom Validity Scale (FBS): Verified utility in forensic practice. *The Clinical Neuropsychologist, 24(4)*, 701-724.

O'Bryant, D. E., & Lucas, J. A. (2006). Estimating the Predictive Value of the Test of Memory and Malingering: An illustrative example for clinicians. *The Clinical Neuropsychologist, 20(3),* 533–540.

Osmon, D. C. & Mano, Q. R. (2009). Malingered attention deficit hyperactivity disorder: Effort, depression, and dependence in the pursuit of academic accommodations. In J. E. Morgan & J. J. Sweet (Eds.) *Neuropsychology of malingering casebook* (pp. 386-397). New York: Psychology Press.

Osmon, D.C., Plambeck, E., Klein, L, & Mano, Q. (2006). The word reading test of effort in adult learning disability: A simulation study. *The Clinical Neuropsychologist, 20(2)*, 315-324.

Petrila, J., & Brink, T. (2001). Mental illness and changing definitions of disability under the Americans with Disability Act. *Psychiatric Services, 52(2),* 626-630.

Quinn, C. A. (2003). Detection of malingering in assessment of adult ADHD. *Archives of Clinical Neuropsychology, 18*, 379-395.

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

Rogers, R. (2008). *Clinical assessment of malingering and deception*. New York: Guilford Press.

Rogers, R., Gillis, J.R., Dickens, S.E., & Bagby, R.M. (1991). Standardized assessment of malingering: Validation of the Structured Interview of Reported Symptoms. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3(1)*, 89-96.

Rogers, R., Harrell, E. H., & Liff, C. D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review, 13*, 255-274.

Rohling, M. L. & Demakis, G. J. (2010). Bowden, Shores, & Mathias (2006): Failure to replicate or just failure to notice. Does effort still account for more variance in neuropsychological test scores than TBI severity? *The Clinical Neuropsychologist,24(1)*, 119-136.

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

Simon, M. J. (2007). Performance of mentally retarded forensic patients on the Test of Memory Malingering. *Journal of Clinical Psychology, 63(4),* 339-344.

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.

Smith, G.P. & Burger, G.K. (1997). Detection of malingering: Validation of the Structured Inventory of Malingered Symptomology (SIMS). *Journal of the American Academy of Psychiatry and the Law, 25(2)*, 183-189.

Suhr, J., Hammers, D., Dobbins-Buckland, K., Zimak, E., & Hughes, C. (2008). The relationship of malingering test failure to self-reported symptoms and neuropsychological findings in adults referred for ADHD evaluation. *Archives of Clinical Neuropsychology, 23*, 521-553.

Sullivan, B. K., May, K., & Galbally, L. (2007). Symptom exaggeration by college adults in Attention-Deficit Hyperactivity Disorder and Learning Disorder assessments. *Applied Neuropsychology, 14(3),* 189-207.

Sullivan, K. & King, J. (2010). Detecting faked psychopathology: A comparison of two tests to detect malingered psychopathology using a simulation design. *Psychiatry Research, 176(1)*, 75-81.

Sweet, J. J. (2009). Neuropsychology and the law: Malingering assessment in perspective. In J. E. Morgan & J. J. Sweet (Eds.) *Neuropsychology of malingering casebook* (pp. 3-8). New York: Psychology Press

Tombaugh, T. N. (1996). *Test of Memory Malingering: TOMM*. New York: Multi-Health Systems Inc.

Weinborn, M., Orr, T., Woods, S. P., Conover, E., & Feix, J. (2003). A validation of the Test of Memory Malingering in a forensic psychiatric setting. *Journal of Clinical and Experimental Neuropsychology, 25(7),* 979-990.

White, B. P., Becker-Blease, K. A., & Grace-Bishop, K. (2006). Stimulant medication use, misuse, and abuse in an undergraduate and graduate student sample. *Journal of American College Health, 54(5*), 261-268.

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

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