An item response theory analysis of sex differences with the Miller Forensic Assessment of Symptoms Test

Megan Thomet
Pacific University

Recommended Citation

This Dissertation is brought to you for free and open access by the Theses, Dissertations and Capstone Projects at CommonKnowledge. It has been accepted for inclusion in School of Professional Psychology by an authorized administrator of CommonKnowledge. For more information, please contact CommonKnowledge@pacificu.edu.
An item response theory analysis of sex differences with the Miller Forensic Assessment of Symptoms Test

Abstract
Malingering is present in all settings; however, rates of malingering are higher in forensic settings than in the community. Males make up the vast majority of the incarcerated population and therefore the psychological measures created for this population have focused on males. With the rapidly increasing rate of females entering the correctional system, it is important to assess the utility of these measures with both males and females. The Miller Forensic Assessment of Symptoms Test (M-FAST) is a screening measure for feigning severe mental health symptoms and is often used in forensic and correctional settings. Several M-FAST items assess mood disorder symptoms and there is a known sex difference in the prevalence of mood symptoms. Using a sample of male and female prison inmates, several statistical analyses were conducted to determine whether the mood items function differently between the sexes. The results of the analysis indicated that the mood items functioned similarly with both males and females and that females were no more likely to endorse the mood items than males. These results are discussed in terms of use of the M-FAST with both males and females in correctional settings.

Degree Type
Dissertation

Rights
Terms of use for work posted in CommonKnowledge.

Comments
Library Use: LIH

This dissertation is available at CommonKnowledge: http://commons.pacificu.edu/spp/1132
Copyright and terms of use

If you have downloaded this document directly from the web or from CommonKnowledge, see the “Rights” section on the previous page for the terms of use.

If you have received this document through an interlibrary loan/document delivery service, the following terms of use apply:

Copyright in this work is held by the author(s). You may download or print any portion of this document for personal use only, or for any use that is allowed by fair use (Title 17, §107 U.S.C.). Except for personal or fair use, you or your borrowing library may not reproduce, remix, republish, post, transmit, or distribute this document, or any portion thereof, without the permission of the copyright owner. [Note: If this document is licensed under a Creative Commons license (see “Rights” on the previous page) which allows broader usage rights, your use is governed by the terms of that license.]

Inquiries regarding further use of these materials should be addressed to: CommonKnowledge Rights, Pacific University Library, 2043 College Way, Forest Grove, OR 97116, (503) 352-7209. Email inquiries may be directed to: copyright@pacificu.edu

This dissertation is available at CommonKnowledge: http://commons.pacificu.edu/spp/1132
AN ITEM RESPONSE THEORY ANALYSIS OF SEX DIFFERENCES WITH THE MILLER
FORENSIC ASSESSMENT OF SYMPTOMS TEST

A DISSERTATION
SUBMITTED TO THE FACULTY
OF
SCHOOL OF PROFESSIONAL PSYCHOLOGY
PACIFIC UNIVERSITY
HILLSBORO, OREGON

BY
MEGAN THOMET

IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PSYCHOLOGY

JULY 24, 2014

APPROVED BY THE COMMITTEE:

Michelle R. Guyton, Ph.D., ABPP
Genevieve Arnaut, Ph.D., Psy.D.
Ronna J. Dillinger, Ph.D., ABPP
Abstract
Malingering is present in all settings; however, rates of malingering are higher in forensic settings than in the community. Males make up the vast majority of the incarcerated population and therefore the psychological measures created for this population have focused on males. With the rapidly increasing rate of females entering the correctional system, it is important to assess the utility of these measures with both males and females. The Miller Forensic Assessment of Symptoms Test (M-FAST) is a screening measure for feigning severe mental health symptoms and is often used in forensic and correctional settings. Several M-FAST items assess mood disorder symptoms and there is a known sex difference in the prevalence of mood symptoms. Using a sample of male and female prison inmates, several statistical analyses were conducted to determine whether the mood items function differently between the sexes. The results of the analysis indicated that the mood items functioned similarly with both males and females and that females were no more likely to endorse the mood items than males. These results are discussed in terms of use of the M-FAST with both males and females in correctional settings.

Keywords: feigning, malingering, Differential Item Functioning (DIF), Item Response Theory (IRT), Miller Forensic Assessment of Symptoms Test (M-FAST)
Acknowledgements

This project was made possible by the generous nature of David Hill, Lea Laffoon, and Orbelin Montes for the use of their databases. I would also like to thank Jonathan Ryan for his assistance with data collection. Lastly, I would like to thank Dr. Guyton for being my dissertation chair, Dr. Arnaut for being my dissertation reader, and Dr. Dillinger for being my dissertation consultant. I am very thankful for the feedback and support my committee provided me both on my dissertation and in my doctoral program over the years.
TABLE OF CONTENTS

ABSTRACT .............................................................................................................. ii

ACKNOWLEDGEMENTS ..................................................................................... iii

LIST OF TABLES ................................................................................................. vii

LIST OF FIGURES ............................................................................................... viii

INTRODUCTION .................................................................................................. 1

LITERATURE REVIEW ....................................................................................... 5
  Malingering Detection Strategies ................................................................. 5
  Malingering Assessment .................................................................................. 8
    Structured Interview of Reported Symptoms-Second Edition ................. 9
    Structured Inventory of Malingered Symptomatology ............................ 14
    Miller Forensic Assessment of Symptoms Test ......................................... 15
  Sex Differences in Assessment ....................................................................... 20
    Assessment of Depressed Mood ............................................................... 22
    Assessment of Symptom Feigning .............................................................. 25
    IRT and the M-FAST .................................................................................... 26

PURPOSE OF THE PRESENT STUDY ................................................................. 30
  Hypothesis ....................................................................................................... 31

METHOD ............................................................................................................. 32
  Participants ...................................................................................................... 32
    Sample 1 ..................................................................................................... 32
    Sample 2 ..................................................................................................... 32
    Honest Condition ....................................................................................... 33

iv
LIST OF TABLES

Table 1. Demographic Characteristics...............................................................37
Table 2. Frequency of Endorsement of M-FAST Items per Category by Sex.............52
Table 3. Frequency of M-FAST Total Score per Category by Sex.............................54
Table 4. Results for 2PL Models with Anchored Items per Sex...............................56
Table 5. Results for DIF Statistics for Items 2, 5, and 23......................................63
Table 6. Chi-Square Test Results for Items 2, 5, and 23.........................................63
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Graph of an Item Characteristic Curve</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>ICCs for Item Difficulty</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>ICCs for Item Discrimination</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>Information Function Curves</td>
<td>46</td>
</tr>
<tr>
<td>5</td>
<td>ICC for Item 2 for Both Groups</td>
<td>57</td>
</tr>
<tr>
<td>6</td>
<td>ICC for Item 5 for Both Groups</td>
<td>58</td>
</tr>
<tr>
<td>7</td>
<td>ICC for Item 23 for Both Groups</td>
<td>59</td>
</tr>
<tr>
<td>8</td>
<td>IFC for Item 2 for Both Groups</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>IFC for Item 5 for Both Groups</td>
<td>61</td>
</tr>
<tr>
<td>10</td>
<td>IFC for Item 23 for Both Groups</td>
<td>62</td>
</tr>
</tbody>
</table>
**Introduction**

The term “maligner” is derived from the French word malingerer, meaning “‘to suffer,’ and also ‘pretend to be ill’” (Online Etymology Dictionary, 2010, para 1). Adaptation of the term in American culture has led to the dissolution of the first definition and intense focus on the concept of symptom feigning. Malingering has been defined 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” (American Psychiatric Association [APA], 2013, p. 726. Although malingering is present in all settings, forensic populations (e.g., inmates, forensic psychiatric patients, etc.) have an increased base rate of malingering that has been estimated at 15 -17% whereas clinical patients have an estimated base rate of 5 -7% (Gillard & Rogers, 2010). Due to the limited resources available to individuals with genuine mental illness in forensic settings, there has been an increase in the number of malingering evaluations conducted (Scott, 2009). There is a need for clinicians to accurately identify individuals who engage in malingering so that available resources are reserved for individuals who truly need them.

Malingering is difficult to assess because measures cannot capture secondary gain, or the person’s intention when reporting symptoms. Therefore, the measures used in malingering evaluations assess response style. Response style is simply “the tendency of a respondent to answer in a specific way regardless of how a question is asked” (Bureau of Justice Assistance, Retrieved February 17, 2012, par 25). The specific response style that is consistent with malingering is symptom feigning. Symptom feigning is the creation or exaggeration of symptoms, problems, or negative characteristics. Typically, evaluations of response style
integrate information from a clinical interview, one or more structured assessments, and collateral sources of data (i.e., medical files, family members). The type of assessments used in these evaluations have evolved, becoming more informed by research on strategies commonly used by individuals who engage in malingering. These include the Structured Interview of Reported Symptoms-Second Edition (SIRS-2; Rogers, Sewell, & Gillard, 2010), the Structured Inventory of Malingered Symptomatology (SIMS; Smith & Burger, 1997), and the Miller Forensic Assessment of Symptoms Test (M-FAST; Miller, 2001). Although these instruments are currently used for assessing symptom feigning, they are imperfect. Part of their unreliability is due to limitations of the theory used to create them. As with most assessments in psychology, these feigning assessments are based on Classical Test Theory (CTT; Harvey & Hammer, 1999). CTT, also known as a True Score Theory, holds that every observed score is comprised of true ability and random error (Kline, 2005). Although CTT has been used to develop assessments tools, there are also certain inherent limitations. These limitations include the sole use of total scores or scale scores and the lasting influence of the norming or validation sample (Emberton & Reise, 2000).

In order to compensate for some of the limitations of CTT, psychologists have begun using Item Response Theory (IRT) to construct and evaluate psychological measures. IRT is a latent trait theoretical model, and the premise of the theory is that an examinee’s performance on an item can be predicted by a set of factors called latent traits, which is denoted with the Greek symbol $\theta$ (Hambleton, Swaminathan, & Rogers, 1991). Unlike CTT, IRT focuses on individual items when predicting where an examinee falls along the latent trait. The greatest difference between the theories is that IRT values are invariant; the score obtained on the item is not influenced by the validation sample and is a true representation of where an individual falls
along the latent trait. In contrast, when using CTT, the examinee’s score is influenced by many factors, such as the norming population (Kline, 2005). In CTT terms, the score obtained is the true score, which eliminates the problem of error introduced during the assessment process.

The ability of IRT to identify where an individual falls along the latent trait allows researchers to evaluate how different groups of individuals score on items and whether different groups systematically score differently. With this tool, researchers can test the generalizability of items across ethnic groups or between men and women. For example, there is ample research on sex differences in the prevalence and expression of mood symptoms. Research has shown that females tend to have higher scores than do males on measures of mood symptoms, but this finding has not always been consistent (Santor, Ramsay, & Zuroff, 1994). Analyses of individual items on some of these measures have shown that females and males systematically score differently on items (e.g., Santor et al., 1994).

With increasing numbers of females entering the legal system, it is becoming more necessary to understand how current assessments for symptom feigning can be generalized across males and females. According to the Bureau of Justice (2012), there was a 20% increase of female inmates in prisons in the United States from 2000 to 2009 compared to a 14% increase for males. In 2009 women represented 18% of the total prison population (Glaze, 2009). For females, the types of crimes committed has changed with a trend towards serious crimes and more drug-related offenses due to the war on drugs (Lewis, 2009). Females have been incarcerated at a much lower rate than males, but rates of incarceration for females are growing disproportionately to those of males. Thus, it is important that forensic assessment measures, which have typically been constructed and validated with males, are also validated with female populations. Information about sex differences in mental health must be applied to forensic
assessment to make sure that clinical and legal decisions made about women are reliable and accurate. One area of mental health that is known to demonstrate large differences between males and females is in the expression of mood symptoms, specifically depressive symptoms.

The purpose of the current study was to conduct an IRT analysis of the M-FAST to assess whether items on the M-FAST show sex bias. The M-FAST is a commonly used measure to screen for malingering in forensic settings; it has four items that address mood symptoms (Miller, 2001). Therefore, I hypothesized that an IRT analysis of the items would reveal whether females systematically score differently on the mood symptom items than do males.
Literature Review

Malingering is most prevalent in forensic settings where clinicians have estimated that 1 in 6 examinees are suspected of malingering (Green & Rosenfeld, 2011). Incentives to malinger that are frequently found in correctional settings include obtaining prescription medication for consumption or sale, obtaining a transfer to a psychiatric hospital with the intention of escaping or doing “easier time,” or obtaining a cell transfer to a more desirable location (Guy & Miller, 2004; Resnick, 1999, p. 160). An examinee’s perceptions of a forensic setting can increase a client’s desire to deceive (Vitacco & Rogers, 2009). An example would be a situation in which the examinee perceives the clinician to be working for the agency, not for the examinee. This may happen in evaluations in which a clinician is hired by the court or in correctional settings in which the clinician works for the institution. If the examinee believes the clinician is looking for the best outcome for the courts or the correctional facility, he or she may feel inclined to exaggerate symptoms to ensure he or she obtains the sought-after resources.

Detection of malingering has become a larger concern as the number of incarcerated mentally ill individuals has steadily increased, and correctional environments are becoming the largest providers of mental health services (Scott, 2009). Because of the large numbers of inmates with mental illness and the limited resources of the prison environment, correctional administrations need to identify who is truly in need of services. Individuals who feign mental illness are costly to the correctional system. If resources are used on individuals who are not in need there may be a denial of services to other individuals who truly need the services.

Malingering Detection Strategies

Malingering by definition is not a stable attribute of a person but instead is a dynamic factor that is influenced by the circumstances and consequences of the person’s situation.
(Conroy & Kwartner, 2006). Likewise, there is no known biological basis, etiology, symptom course, or treatment intervention for malingering. Rogers described malingering as an adaptational process where individuals utilize a cost-benefit analysis to systematically determine whether this response style is appropriate for the circumstance (Rogers, 1990; Rogers, 2008b).

Detection strategies for malingering tend to be based on known etiology and symptom course of formal diagnoses. Rogers (2008a) provided a thorough description of detection strategies. Detection strategies can be characterized into two domains: unlikely presentations and amplified presentations. Unlikely presentation detection strategies can be further broken down into two substrategies. The first substrategy is based on the presence of symptoms that are rarely endorsed by genuine patients. The second substrategy is based on the absence of symptoms generally endorsed by genuine patients. Individuals who engage in malingering are often unaware of symptoms that are rare or unusual in genuine pathology and therefore may endorse them at higher rates. This strategy is one of the most robust detection strategies (Conroy & Kwartner, 2006; Gillard & Rogers, 2010). Improbable symptoms are different from rare and unusual symptoms because they have an outrageous or preposterous quality to them and would rarely if ever be endorsed by genuine patients. Individuals who engage in malingering may be knowledgeable about common symptoms. In an effort to malinger, they then endorse these common symptoms simultaneously, without knowing that such a combination is very rare in clinical populations. This detection strategy, called symptom combination, works on the presumption that few non clinicians know which symptoms tend to co-occur to create different disorders.

Rogers (2008a) also described amplified presentations. Amplified presentations are based on the frequency and intensity of reported symptoms. Individuals who engage in malingering
may endorse symptoms across multiple diagnostic domains, sometimes called the “more is better” or *indiscriminant symptoms* strategy. Other individuals may exaggerate the intensity and *severity* of the reported symptoms. The *obvious versus subtle* symptom detection strategy is related to the common symptoms previously mentioned but takes into consideration the frequency of the reported symptoms. Diagnoses tend to have obvious symptoms that characterize the pathology for lay people as well as subtle symptoms that are less commonly known among the lay population. Individuals who engage in malingering tend to endorse more obvious and less subtle symptoms than do individuals with genuine pathology. Using another detection strategy, *erroneous stereotypes*, the evaluator assesses the use of common misconceptions and stereotypes about mental illnesses that may be used when an individual feigns symptoms. An example would be describing a person with schizophrenia as “having two personalities” (Conroy & Kwartner, 2006, p. 38). The last strategy, *reported versus observed symptoms*, combines behavioral observations and reported symptoms. Quite simply, this strategy is used to evaluate inconsistencies between the examinee’s actual observed behavior and reported symptoms. These are some of the typical detection strategies frequently employed in creating items for assessments of response style (Conroy & Kwartner, 2006; Gillard & Rogers, 2010; Rogers, 2008a; Rogers & Bender, 2003). Although these detection strategies are helpful, if an evaluator holds erroneous stereotypes and misconceptions about malingering, the results can have negative consequences.

Rogers and Bender (2003) highlighted a few misconceptions about malingering that an examiner needs to be aware of when conducting assessments. The first erroneous assumption is that malingering and mental illness are mutually exclusive concepts. There is nothing inherent in either malingering or mental disorders that rules out the possibility of the other. In fact,
individuals with mental disorders can be more effective at malingering and can exaggerate current symptoms (Conroy & Kwartner, 2006). Similarly, an individual who desists from malingering may still have underlying mental health issues. This means that feigning mental illness can be both a production of symptoms and exaggeration of symptoms. There are other misconceptions that malingering is either very rare or very common. Research has shown neither of these to be true with base rates in forensic cases being 15 -17% (Gillard & Rogers, 2010; Rogers & Bender, 2003).

**Malingering Assessment**

The strategies and approaches used to detect response styles have evolved with changes in the field such as changes in legal consequences for different diagnoses. The strategy for detecting response styles that has been employed the longest is unstructured clinical judgment (Rogers & Bender, 2003). The basis of clinical judgment is that psychologists can detect response styles based on clinical acumen alone (Rogers & Bender, 2003). Unfortunately, this strategy lacks empirical support. Ekman and O’Sullivan (1991) studied the accuracy of multiple groups who had professional interests in lying (e.g., U.S. Secret Service, judges, psychiatrists) to accurately identify people who were lying. They found that psychiatrists’ abilities to detect lying were not significantly different from chance. Furthermore, DePaulo and Pfeifer (1986) found no relationship between a person’s amount of experience or confidence in their ability to detect lying and their actual ability to detect lying. In other words, being confident in one’s lie-detecting abilities did not influence accuracy in detecting false information.

In order to address the lack of empirical support for clinical judgment, evaluators began using already established measures to assist in diagnosing malingering. This strategy is convenient because there are empirically supported multiscale inventories that devote at least one
scale to measuring response style (Rogers & Bender, 2003). The Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Williams, Graham, Tellegen, & Kaemmer, 1989) and the Personality Assessment Inventory (PAI; Morey, 1991) are widely used broadband measures of personality and psychopathology that also have multiple scales to detect symptom feigning (Gillard & Rogers, 2010).

In the early 1990s there was a move to create measures specifically for detecting response style in order to increase classification accuracy (Rogers & Bender, 2003). Although classification accuracy is important for all diagnoses, due to the potential legal consequences and denial of services to individuals labeled malingerers, there is a need for higher levels of certainty. The Structured Interview of Reported Symptoms (SIRS) was created to further minimize false-positives and uses eight of the detection strategies previously outlined (Rogers & Bender, 2003). The second version of the measure was published in 2010 (SIRS-2) and was validated with four samples: a community sample, an outpatient sample, an inpatient forensic sample, and an inpatient sample (Rogers et al., 2010). The SIMS is a screening measure of symptom feigning that was validated on college students. The M-FAST is a self-report screening measure for malingering that can be administered in about 5 min (Miller, 2001). It also utilizes some behavioral observations (Miller, 2001). The M-FAST was validated using both an inpatient forensic sample and a college sample.

**Structured Interview of Reported Symptoms – Second Edition.** The original SIRS was created in 1992 by Rogers, Bagby and Dickens. It was a structured interview composed of 172 items that had an administration time of about 25 min (Rogers et al., 1992). Items are scored X (no information), 0 (not present), 1 (sometimes) or 2 (definitely yes). It contained eight primary scales: Rare symptoms (RS), Symptom Combination (SC), Improbable or Absurd (IA),
Blatant Symptoms (BS), Subtle Symptoms (SU), Selectivity of Symptoms (SEL), Severity of Symptoms (SEV), and Reported versus Observed (R). It consisted of five supplementary scales: Direct Appraisal of Honest (DA), Defensive Symptoms (DS), Overly Specified Symptoms (OS), Inconsistency of Symptoms (INC) and Self-Management of Symptoms (SM).

In 2010, Rogers, Sewell and Gillard published the Structured Interview of Reported Symptoms- Second Edition (SIRS-2). There are many similarities between the original version and the new version. The measure is still a structured interview composed of 172 items that has an administration time of about 45 min with the same scoring (Rogers et al., 2010). It consists of the same eight primary scales as the original version. The five supplementary scales have been changed, with some of the scales being updated, one being discarded, and one being created. The scales that have been updated are: Direct Appraisal of Honesty (DA), Defensive Symptoms (DS), Overly Specified Symptoms (OS), and Inconsistency of Symptoms (INC). The Self-Management of Symptoms scale has been removed and a new scale called Improbable Failure (IF) has been added. The IF scale is used to screen for the feigning of cognitive symptoms. The items were present in the SIRS; however, due to limited validation studies, it was not recognized as a scale (Rogers et al., 2010).

The major changes in the SIRS-2 predominantly have to do with classification. To determine classification, the SIRS-2 included a decision model that is intended to reduce the number of false positives. The decision model uses some of the same criteria in the SIRS (i.e., one or more scale in the definite range, three or more primary scales in the probable range), however, additional indices and totals that assist the process. The Rare Symptoms-Total (RS-Total) “is specifically intended to differentiate between (a) genuine but atypical presentations and (b) feigned presentations” (Rogers et al., 2010, p. 69) to reduce false positives. The Modified
Total Index (MT Index) has replaced the SIRS Total Score and is used in “differentiating hard-to-classify cases while minimizing false positives” (Rogers et al., 2010, p. 70). The Supplementary Scale Index (SS Index) is used “to identify ‘too-good-to-be-true’ SIRS profiles produced by likely feigners” (Rogers et al., 2010, p. 70).

The categories for the overall classification have also been modified (Rogers et al., 2010). In addition to the Genuine Responding and Feigning categories of the SIRS, now there are also the Indeterminate-Evaluate, Indeterminate-General and Disengagement: Indeterminate-Evaluate categories. When an examinee falls in the Indeterminate-Evaluate category, the authors advise that there can be no determination of feigning, and they recommend further assessment. When an examinee falls in the Indeterminate-General category, the authors once again advise that there can be no determination of feigning and that the evaluator look to collateral data. The Disengagement: Indeterminate-General category is largely influenced by the SS Index and identifies individuals who attempt to hide their feigning by not responding to many items (Rogers, et al, 2010). The utility estimates for the SIRS-2 are as follows: false-positive rate of 2.5%, sensitivity of .80, specificity of .975, positive predictive power (PPP) of .91, negative predictive power of .91, and overall correct classification of .91.

Although the SIRS-2 was meant to replace the SIRS (Rogers, Bagby, & Dickens, 1992) as the gold standard for measuring response style, the limited research that has been published on the measure has highlighted many areas of concern (DeClue, 2011; Rubenzer, 2010). An overarching theme among the articles is the lack of information presented in the manual. It has also been argued that the authors failed to follow both standards agreed upon in the scientific community as well as their own advice. These standards include withholding data for validation studies, failing to adequately identify demographic information about the samples,
misclassification of pertinent information in the manual, and lack of peer review (DeClue, 2011; Rubenzer, 2010). There have been many criticisms concerning the inclusion of a sample of traumatized inpatients with a high rate of dissociative symptoms. Research on the SIRS had shown high false-positive rates for individuals with dissociative symptoms, and there has been concern as to how inclusion of this sample has affected the SIRS-2 (DeClue, 2011).

Because the SIRS was considered the gold standard for feigning assessments, the creators of the SIRS-2 wanted to ensure the large body of research for the SIRS could be generalized to the SIRS-2. The SIRS was validated with many forensic populations, including correctional populations, both psychiatric units and general population, jail populations, and disability compensation populations (Edens, Poythress, & Watkins-Clay, 2007; McDermott & Sokolov, 2009; Rogers, Gillis, & Bagby, 1990; Rogers, Payne, Berry & Granacher, 2009). The reliability of the measure is very good, with alpha coefficients of the primary and supplementary scales ranging from 0.77-0.92 (Rogers, 2008b). Rogers (2008b) summarized studies published after the manual was published and found very similar alpha coefficients. He also found the average interrater reliability to be 0.99. Lastly, he found small standard error of measurements, with the largest being 0.60. This means that evaluators can be confident that the SIRS scores give a reliable estimate of the person’s true score. A factor analysis of the SIRS scores resulted in a robust two-factor loading of unlikely and amplified presentations (Rogers, 2008b). Rogers, Jackson, Sewell and Salekin (2005) conducted a confirmatory factor analysis of the SIRS and confirmed the two-factor model of unlikely and amplified presentations, although they named their factors spurious and plausible presentations. Rogers (2008b) further reviewed the validity of the SIRS. He found that the research showed strong convergent validity between the SIRS and the MMPI-2 feigning scales and the PAI validity scales. As for discriminant validity, the SIRS
discriminated fairly well between individuals who engage in malingering and those responding honestly as evidenced by large effect sizes.

Green and Rosenfeld (2011) conducted a meta-analysis of the SIRS using 26 studies published between 1990 and 2009. Among the many analyses conducted, they compared initial development studies to studies published after the manual was published. They found lower effect sizes in later studies, which means that studies published after the manual was published showed the assessment tool was weaker at distinguishing individuals who engage in malingering from those who respond honestly. There was also a trend for lower specification rates, which is the percentage of honest responders who were correctly classified. Because the SIRS was developed to minimize false positives, this rate is concerning. The opposite was found for sensitivity rates, which increased. This means that the studies published after the SIRS manual was published on average had higher percentages of correctly identifying individuals who were suspected of malingering. After analyzing the effect sizes for all of the studies, Green and Rosenfeld generated a composite effect size of $d = 2.02$ for the total score. This is a very large effect size and indicates that individuals who feign symptoms are likely to endorse a large number of items, and genuine responders are likely to endorse few items. Lastly, they found average effect sizes for studies that compared individuals who engaged in malingering with nonclinical samples as opposed to comparing individuals who engaged in malingering with clinical samples. Additionally, the SIRS was better able to distinguish honest responding in offender and community samples than in clinical samples. Many items on the SIRS are symptoms that are frequently experienced by genuinely mentally ill individuals; therefore, individuals from clinical populations may score higher than honest responding nonclinical samples without producing or exaggerating symptoms.
Structured Inventory of Malingered Symptomatology. As already mentioned, the long administration time of the SIRS has made a niche for brief screening measures of response style. The SIMS was specifically created to have an administration time that is less than the SIRS-2. The test is comprised of 75 true-or-false items in a self-report format (Dunn, 2007; Smith & Burger, 1997). The average administration time is 10-15 min. The SIMS consists of five scales: Psychosis (P), Neurologic Impairment (N), Amnestic Disorders (AM), Low Intelligence (LI) and Affective Disorders (AF). As can be seen from the five scales, the SIMS measures malingering across psychological and cognitive domains, which the M-FAST fails to do. The detection strategies used are identification of rare symptoms, improbable symptoms, and unlikely symptom combination (Smith, 2008). The test is scored by summing the number of critical items endorsed to yield a total score. The initial validation of the measure resulted in interrater reliabilities for the scales ranging from 0.76 (N)-0.95 (AF) and internal consistency reliabilities ranging from 0.80-0.84. Smith and Burger (1997) calculated the sensitivity, specificity and efficiency score. The total score had a sensitivity of 95.6%, a specificity of 87.9% and an efficiency of 94.5%. The scales ranged on sensitivity from 74.6-88.3%, on specificity from 51.5-90.9% and efficiency from 74.4-88.7%. An exploratory factor analysis was conducted which resulted in a five-factor model: however, some N scale items cross-loaded with P scale items (Smith, 2008).

By their nature, screening measures should have high sensitivity and negative predictive power (NPP; Smith, 2008). NPP is the probability that a person whose score suggests honest responding (i.e., not malingering) is truly not malingering (VanDerHeyden & Burns, 2010). On the other hand, positive predictive power (PPP) is the probability that a person whose score suggests malingering is truly malingering (VanDerHeyden & Burns, 2010). Smith (2008)
compiled the research on the sensitivity PPP and NPP of the SIMS. Known-group comparison studies for the SIMS have shown high NPPs ranging from 0.75-1.00. Higher NPP scores were achieved when the total score cutoff was increased to 16. Sensitivity estimates for known-group comparison studies ranged from 0.85-1.00. PPP for the SIMS has tended to be moderate, ranging from 0.28-0.55.

**Miller Forensic Assessment of Symptoms Test (M-FAST).** The M-FAST is a 25-item structured interview screening tool that assesses response style. The M-FAST requires approximately 5 min to administer. This test utilizes seven specific strategies for detection: identification of unusual hallucinations, identification of reported versus observed symptoms, identification of extreme symptomatology, identification of rare combinations, identification of negative image, identification of unusual symptom course, and identification of suggestibility (Miller, 2001). The majority of the items on the measure were created to mimic psychotic symptoms; however four items were created to mimic mood symptoms, specifically depression and mania. Because it is a screening measure, it was designed to yield higher false positives and reduce false negatives. High scores on the M-FAST are not automatically indicative of feigning or exaggeration; instead, high scores indicate a more rigorous assessment should be used to assess whether the individual is feigning.

A statistical analysis technique that is frequently used when looking at prediction is the Receiver Operator Characteristics (ROC) curve. The result of ROC analysis is a graph of a curve. The Area Under the Curve (AUC) is used to determine how well the measure predicts the intended outcome (Mandrekar, 2010). AUC scores range from 0 to 1, with .5 equal to chance prediction. The closer the AUC is to 1, the better the predictive power (Mandrekar, 2010). Research has shown strong predictive accuracy for the total score of the M-FAST (AUC = .92)
that persists when participants are separated by ethnicity (Guy & Miller, 2004; Miller, 2005). Miller (2004) found that when looking at a sample of 50 criminal defendants found incompetent to stand trial, the M-FAST total score AUC was .95 for differentiating malingerers from honest responders. Participants were categorized based on their SIRS scores. The recommend cutoff score of 6 results in a sensitivity of 86% and specificity of 83% (Guy & Miller, 2004). Other studies have showed the sensitivity of the M-FAST to range from 25 - 93%, the specificity to range from 83 - 95%, and the overall classification rate to range from 86 - 87% (Hill, 2009; Miller, 2001). The misclassification rate was 2% false negatives, 12% false positives and an overall misclassification rate of 14% (Miller, 2004). The positive predictive power has been shown to be 43% and the negative predictive power to be 90% (Hill, 2009). For a screening measure, specificity is one of the most important statistical measures, as it reflects the accurate identification of true negatives; for the M-FAST that would be accurate identification of individuals who are honestly responding. Research on the M-FAST has yielded consistently high specificity.

Although the M-FAST has been available for a short period of time, the body of research on the measure has yielded beneficial information. Miller (2004) tested the utility of the M-FAST in a sample of 50 male criminal defendants found incompetent to stand trial. The participants were administered the SIRS to determine two groups: those labeled as malingering and those labeled as honest responders. Twenty-eight percent of the sample were labeled as malingering on the basis of SIRS scores. Each participant’s score on the M-FAST was then analyzed to determine how well the test predicted the appropriate SIRS group. The results showed that participants labeled malingerers scored significantly higher on the total and scale scores of the M-FAST compared to the honest responder group. The participants also completed
the MMPI-2. There were high correlations between the M-FAST total and scale scores with the MMPI-2 fake-bad scale, indicating convergent validity. Also there were negative or nonsignificant correlations between the M-FAST total and scale scores and the MMPI-2 defensiveness indicators indicating discriminant validity. Lastly, the administration time was analyzed. Although the time ranged from 2-13 min for the entire sample, participants in the malingering group took significantly more time to complete the M-FAST than did the honest responding group.

Jackson, Rogers, and Sewell (2005) also looked at the M-FAST with defendants found incompetent to stand trial. There were two different samples in the study. One sample consisted of male and female inmates from a county jail who participated in a simulation in which the participants were instructed or coached on how to take the test. The inmates were further divided into a group coached to feign on the measure (simulators) and a control group. The other sample consisted of male and female defendants found incompetent to stand trial called the competency sample. The competency sample was further divided into a clinical comparison group and a suspected malingering group based on their SIRS scores.

Jackson et al.’s (2005) results showed that both the simulators and the malingerers endorsed significantly more items than did the controls. Simulators and malingerers also scored significantly higher on all subscales and had significantly higher total scores than did the control groups. There was no significant difference between either of the control groups or between the simulator and malingering groups. Overall the malingering group had higher effect sizes than did the simulators. There was no analysis of sex differences in this study.

Guy and Miller (2004) studied the utility of the M-FAST with a sample of 50 male inmates in a maximum-security state prison. The participants were administered the M-FAST as
well as the SIRS. The SIRS was administered to classify participants as malingerers or honest responders. Participants with a total SIRS score of 76 or above and at least two primary scales in the probable range were classified as malingerers. The results showed that malingerers scored significantly higher on the M-FAST than did the honest responding group on the total score as well as on the individual scale scores. The generalizability of the M-FAST was examined by comparing across race and ethnicity. Caucasians, African Americans and Hispanics in the malingering group and the honest responding group performed similarly on the total score. Furthermore, the generalizability of the cutoff score of 6 across race and ethnicity was examined. The M-FAST perfectly classified all six of the Hispanic participants. The African American group had an AUC of .90 and the Caucasian group had an AUC of .93. Across race, the NPP ranged from .83-1.00, the PPP ranged from .67-.75, the specificity ranged from .63-.90, the sensitivity ranged from .86-1.00.

Miller (2005) further analyzed the generalizability of the M-FAST across race and reading level (specifically literate or illiterate). The sample consisted of 50 male defendants found incompetent to stand trial. The M-FAST and the SIRS were administered to the participants. They were then divided into the honest responding group or the malingering group based on their SIRS score. The utility estimates for race were similar to those found by Guy and Miller (2004). There was no significant difference on the total score or scale scores between African Americans and Caucasians across the groups. The Hispanic group was not included in the utility estimates for race analysis due to the small sample size. All three of the Hispanic participants were correctly classified by the M-FAST. The AUC for the African American group was 1.00 and the AUC for the Caucasian group was .86. Across ethnicities, the NPP ranged from .95-1.00, the PPP ranged from .50-.67, the specificity ranged from .71-.90 and the sensitivity
ranged from .67-1.00. Participants were placed in the illiterate group if they were unable to read several of the items on other measures administered for the study (i.e., SIRS and M Test; Beaber, Marston, Michelli & Mills, 1985). There was no significant difference between the literacy status for either the malingering or honest responding groups. The AUC for the literate group was .94 and the AUC for the illiterate group was .92. Across literacy groups the NPP ranged from .96-1.00, the PPP ranged from .50-.69, the specificity was .83, and the sensitivity ranged from .92-1.00, which suggests similar utility.

Guy, Kwartner, and Miller (2006) investigated the ability of the M-FAST to differentiate feigning of psychotic and affective symptoms. This study consisted of two groups of participants: a simulator group consisting of undergraduate students and a clinical participant group consisting of forensic psychiatric patients found incompetent to stand trial, civil psychiatric inpatients, general population prisoners receiving psychiatric services, and disability claimants applying for outpatient psychiatric services. The simulators were given instructions to feign Schizophrenia, Posttraumatic Stress Disorder (PTSD), Bipolar Disorder, or Depression. The results indicated that the simulators scored significantly higher on the total score across all diagnostic categories than did the clinical participants. The magnitude of the difference was greater in the Schizophrenia and Bipolar groups than in the Depression and PTSD groups. This showed that the simulators were more likely to be identified as malingering than were the clinical participants, especially if they were trying to malinger Schizophrenia or Bipolar Disorder.

The examination of malingering of PTSD has been further researched by different authors. Guriel-Tennant and Fremouw (2006) used a sample of undergraduate students who were all instructed to feign PTSD. They were divided into two groups based on the presence or
absence of a previous traumatic experience and then randomly assigned to either the coached or
the naïve group, resulting in four groups of relatively equal size: trauma positive coached, trauma
negative coached, trauma positive naïve, and trauma negative naïve. The authors opined that a
previous traumatic experience may make people more successful at malingering. All participants
were instructed to feign PTSD symptomatology, with the coached sample being specifically
instructed on psychological test-taking strategies, PTSD diagnostic criteria, and strategies for
avoiding detection. The results showed that all groups had scores above the M-FAST cutoff
score, with participants in both coached groups having lower scores than participants in both
naïve groups. Messer and Fremouw (2007) also looked at malingering of PTSD using an
undergraduate sample divided into a clinical PTSD group, a subclinical PTSD group, a coached
malingering group, and an honest control group. The malingering group scored significantly
higher on the M-FAST than did the other three groups.

Overall, research on malingering measures has covered many domains, such as reliability
and validity across different races and literacy levels. However, one major area that is
significantly lacking is sex differences. It is common knowledge among psychologists that sex
affects the presentation of symptoms and pathology; for example, there is a section in the
Diagnostic and Statistical Manual of Mental Disorder, Fourth Edition, Text Revision (DSM-IV-
TR, APA, 2000) devoted to sex differences for every diagnosis. Yet there is no research about
how females and males might differ in their approach to malingering assessments, especially the
M-FAST.

Sex Differences in Assessment

It has been well-documented that females are more likely than males to be depressed
(APA, 2000; Frank, 2000). The prevalence of depression for adults ranges from 5-9% for
females and 2-3% for males. In a prison population, Fazel and Danesh (2002) found that the rate of depression was higher for both sexes compared to the general population, although rates of depression for females were still higher than for males (12% and 10%, respectively). However, research has shown that sex differences on depression scales are not statistically significant (Nykiel, 2007). One possible explanation is that men and women express depression differently: Men display more somatic-vegetative symptoms and women express more cognitive-affective symptoms (Dozois, Dobson, & Ahnber, 1998; Steer, Beck, & Brown, 1989).

Research on depression among ethnic minority individuals has been inconsistent. Bracken and Reintjes (2010) highlighted that research has been published showing both higher and lower rates of depression among Hispanics compared to Caucasians across the lifespan. However, research suggests that higher rates of depression among females than among males tend to remain consistent across ethnicity, specifically Hispanic populations (Carmody, 2005; Kuehner, 2003).

There is less research on gender differences and Bipolar Disorders, specifically manic symptoms as compared to depressive symptoms. Major Depressive Episodes are predominant for females over Manic and Hypomanic Episodes, whereas for males, the number of Manic or Hypomanic Episodes tend to equal or exceed Major Depressive Episodes (APA, 2013; Viguera, Baldessarini, & Tondo, 2001). For females mixed or depressive symptoms are more common during Manic episodes as opposed to only manic symptoms which are more commonly experienced by males (APA, 2013; Viguera, Baldessarini, & Tondo, 2001). However, there is conflicting research that has suggested no gender differences in symptom presentation of Bipolar Disorders (APA, 2013; Suominen et al., 2009).
Assessment of depressed mood. Before reviewing literature on mood assessments, it will be beneficial to define certain IRT vernacular. IRT is operationally defined in the Method section of this paper. As previously mentioned, IRT is a latent trait theoretical model (Hambleton, et al., 1991). A latent trait is the construct being measured (e.g., depression or feigning) and is denoted by the symbol $\theta$ (Hambleton et al., 1991). There are two parameters that are used to evaluate the properties of an item on a measure: difficulty and discrimination (Baker, 2001). The difficulty of an item represents where that item falls along the latent trait continuum; in other words, does it fall on the high end or the low end (Baker, 2001). Discrimination describes how well the item differentiates between people who score above and those who score below the difficulty point (Baker, 2001). The concept of invariance is integral to IRT. Invariance is essentially measuring item bias (Santor et al., 1994). If an item is invariant then the item is free from bias, whereas if an item is noninvariant then it is biased against a group (Santor et al., 1994). Lastly, when evaluating a measure using IRT, a model of best fit is used to analyze the data (Baker, 2001). There are three models: one-parameter model, two-parameter model, and three-parameter model (Baker, 2001).

Santor et al. (1994) conducted an IRT analysis on the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) with a college sample and a sample of adult outpatients diagnosed with depression. Specifically, the authors were looking for sex bias on the item level, as well as bias in the response options (on the BDI, each item has four response options). According to the authors, “item bias is observed when individuals who are equally depressed, that is who are at the same point on the depression continuum, $\theta$, endorse items or options differently” (p. 256). They found three items that showed item bias: Item 14 (distortion of body), Item 6 (sense of punishment), and Item 10 (crying). Item 14 had the highest bias, with
females responding more strongly to the options than males, reflecting more severe depression at all levels of depression compared to males. Analysis of the options showed that the highest bias was in the central range of the depression continuum. For Item 6, the bias was reversed in that males reported higher levels than females. For Item 10, the bias was most apparent for Option 0 and in the central range of the depression continuum.

In 1994, the *Diagnostic and Statistical Manual, Fourth Edition (DMS-IV; APA, 1994)* was released, in which the criteria for Major Depressive Disorder were changed. Uebelacker, Strong, Weinstock, and Miller (2009) conducted an Item Response Theory Differential Item Functioning (DIF) analysis, which is a measure of item bias, on the new criteria. It should be noted that the diagnostic criteria for Major Depressive Disorder have remained unchanged with the introduction of the *DSM-5* (APA, 2013). The data were taken from the National Epidemiological Survey on Alcohol and Related Conditions, which consisted of a nationally representative sample from the United States of America. Only people who endorsed either depressed mood or anhedonia completed the section on *DSM-IV* mood disorder symptoms. This subsample consisted of adult males and females. The analysis resulted in a clinically and statistically significant item bias for severity of the depression on the appetite/weight disturbance and the fatigue items. This result indicated that, for the same level of depression severity, females tended to be more likely to endorse both of these items than were males. The appetite/weight item was also statistically significant for the discrimination parameter, along with concentration difficulties and suicide, with females endorsing these items more than males.

Osman, Kooper, Barrios, Gutierrez, and Bagge (2004) looked for invariance across sex on items of the Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996), which is based on the *DSM-IV* criteria for depression. If an item has invariance it is essentially measuring
the same underlying construct for all groups. If an item is noninvariant, then it has bias. The authors used an adolescent psychiatric inpatient sample, consisting of males and females, ages 13-17 years. They found three items with noninvariance: Items 7 (self-dislike), 8 (self-criticism) and 18 (changes in appetite). A confirmatory factor analysis revealed a two-factor model, with Factor 1 being a cognitive-affective factor, and Factor 2 being a somatic factor. Items 7 and 8 loaded on Factor 1, and Item 18 loaded on Factor 2. Other researchers measured latent means between male and female adolescents on the Reynolds Adolescent Depression Scale (Reynolds, 1987; Fonseca-Pedrero et al., 2010). The authors found that the females obtained statistically significantly higher scores than did males on items measuring somatic complaints, negative self-evaluation, and dysphoric mood items. Males in turn, scored significantly higher than did females on items measuring anhedonia.

Wu (2010) measured invariance and latent mean differences across sex using the Chinese version of the BDI-II with a sample of college students. A confirmatory factor analysis resulted in a three-factor model of negative attitude, somatic element, and performance difficulty. He found a significant latent mean difference on the negative attitude factor with females endorsing higher scores. He also found five noninvariant items (2, 3, 4, 7, and 10), the majority of which are on the negative attitude scale. Of the noninvariant items, Item 7 (self-dislike) and Item 10 (crying) overrepresented females, and Items 2 (pessimism), 3 (failure), and 4 (loss of pleasure), overrepresented males. Because there has been consistent research showing noninvariance on items due to sex on measures of depression, it is possible that the mood items on the M-FAST may show noninvariance.

Few measures were designed to specifically assess Bipolar Disorder. Per Miller, Johnson and Eisner (2009), the Structured Clinical Inventory for DSM-IV (SCID; First, Spitzer, Gibbon,
and Williams, 1997) is one of the most commonly used assessments to diagnose Bipolar Disorders. Because the assessment is a structured interview, an IRT analysis cannot be conducted. Miller et al. (2009) identified multiple self-report measures that are also utilized in the diagnosis of Bipolar Disorders; however, no research was located in which an IRT analysis was conducted on the measure.

Assessment of symptom feigning. Sex differences have been largely ignored in research on malingering assessment. The majority of samples used in studies of malingering have been comprised of males. Research that does include female participants has tended to employ undergraduate participants and the effect of sex was not assessed. The M Test, a self-report malingering measure, actually has two items that are specific to males (Smith, 2008). It has been a trend in psychology to attempt to adapt already established and widely accepted theories and assessments to different populations, as seen in the downward extension of adult constructs and theories to juveniles, the generalization of measures normed on a Caucasian population to a wide range of ethnicities, and the generalization of theories and measures used with males to females.

More recently, researchers have recognized the error of assuming these generalizations and have been researching theories, constructs, and measures with different populations. Unfortunately, there has been little research on sex differences and issues of response style. The M-FAST is comprised of four affective items: Item 2 (I feel depressed most of the time), Item 3 (Some days I have major mood swings, where for a while I feel great and then I feel depressed), Item 5 (I feel unusually happy most of the time), and Item 23 (Most of the time I feel that I don’t really matter). Although Item 3 is a question related to mood symptoms, this item is different from the other three mentioned as it includes a second part. If the examinee endorses this item, a follow-up question is asked; Does this only happen when you believe that someone is after you?
The answer to the follow-up question is then recorded, not the answer to the mood portion of the question. Smith (2008) located only one study in which the authors analyzed sex differences on the M-FAST, and the researchers found that there was no significant difference for the total score. Because females tend to have higher rates of depression than males and are thought to express their depression with affective symptoms, it is possible that females would systematically approach affective symptoms on a malingering assessment differently than males. However, there has been no research conducted on sex differences among subscales or psychotic versus affective questions. Systematic differences between sexes on the affective items could result in inflated scores for females on the M-FAST.

Traditionally, researchers have been satisfied to evaluate sex differences at the total score level (e.g., Osman et al., 1997; Steer & Clark, 1997). If they failed to find significant differences, that was seen as sufficient to qualify the test as effective with both sexes. Part of the reason for failing to analyze sex differences on individual items was the lack of efficient analytic strategies. The majority of psychological measures, including the M-FAST, have been developed and analyzed using Classical Test Theory which revolves around a measure’s total score (Baker, 2001).

**IRT and the M-FAST.** As psychology progresses, there has been a push to create new assessments using IRT methodology instead of CTT. At the same time, measures frequently used in the field that were not created with IRT are being analyzed with IRT to glean the beneficial information that is not available with CTT. The M-FAST (Miller, 2001) is relatively new, so the research available for the measure is limited.

Researchers have studied the latent construct of the M-FAST. Miller (2001) performed an exploratory factor analysis and concluded that the M-FAST had one prevailing factor
“representative of response styles indicative of malingering” (p. 30) that accounted for 55% of the variance. Vitacco et al. (2008) performed a confirmatory factor analysis which confirmed the one-factor model and showed that 18 out of the 25 items had large thresholds, meaning they were very good at detecting malingering. The three items with the lowest thresholds were Items 2, 5, and 23, all of which are mood items. They also conducted a latent model comparison between the M-FAST and the SIRS. The SIRS has been found to have two factors: *spurious presentation*, which represents detection strategies that use unusual or inconsistent symptoms, and *plausible presentation*, which represents detection strategies that use magnitude of symptoms (Rogers et al., 2005). Although the M-FAST was correlated with both factors of the SIRS, it was more strongly correlated with the spurious presentation ($r = .75$) than the plausible presentation ($r = .61$). The authors concluded that these findings demonstrated good construct validity for the M-FAST.

Rinaldo (2005) studied the M-FAST using IRT, with three goals for the study. The first goal was to evaluate whether the M-FAST met the basic assumptions of IRT, specifically dimensionality, appropriate number of parameters, and monotonicity. The second goal, contingent upon the basic assumptions of IRT being met via the first goal, was to describe the item parameters and scale characteristics of the M-FAST. The last goal was to assess the fit of the selected IRT model in describing the data. The goal of dimensionality is to establish that the measure consists of a single construct, which for the M-FAST is response style (DeMars, 2010). Fit is established to determine which model is the best fit for the items (DeMars, 2010). As previously mentioned, the probability of endorsing an item is a function of the item parameters; therefore, it is necessary to calculate the best-fit model (Harvey & Hammer, 1999). Monotonicity
is the assumption that, as the raw score increases, the probability of endorsing the item increases as well (Junker & Sijtsma, 2000).

The sample consisted of 602 undergraduate students who were predominately Caucasian females, although he did not look at sex differences on individual items. The students were given instructions to feign mental illness. There were three groups: mild, moderate and severe. Since this was the first known study to utilize IRT with the M-FAST, Rinaldo (2005) began by establishing that the M-FAST met the assumptions of IRT. He found that the M-FAST had one dominant first factor, which was consistent with the confirmatory factor analysis conducted by Vitacco et al., (2008). Four of the 25 items were a poor fit in the first factor (Items 2, 5, 11, and 23) and the majority were mood-related questions. Three of the items are the same items Vitacco et al. (2008) identified as being problematic (Items 2, 5, and 23). A second analysis was conducted without the four problematic items. The results indicated a slightly better fit; however, the change after the items were removed was small. The second assumption of IRT is monotonicity, which essentially means that the probability of endorsing items should increase systematically as the total raw score increases. The majority of the M-FAST items conformed to this assumption. The four previously mentioned problematic items, when plotted, showed erratic increases, and in some cases decreased, as the total raw score increased.

Rinaldo’s (2005) second goal was to establish the proper IRT parameter model. All three parameter models were evaluated. Due to the elevated endorsement of Items 2, 5, and 23 by the honest-responding group as well as the previously mentioned violations of assumptions, these items were eliminated in the analysis of parameter model fit. The 2-parameter model was found to be the best fit for the M-FAST. The specific item parameters were then calculated for each item, including the problematic items; however, because the items did not conform well to IRT
they were not recommended for interpretation. Of the remaining items, Rinaldo found the item difficulty (b) to range from -0.91-1.21. The majority of the total scores for the sample fell within -1 to 1, so the difficulty parameter appeared to be most sensitive where the majority of the scores fell. The discrimination (a) ranged from 0.39-1.46, indicating that small differences resulted in different responses on the items. Half of the remaining items had good discrimination (slope of 1.0 or above). He also evaluated the maximum information contributed by each item. Each item had information for every level of the latent trait. The four problematic items all had maximum information scores of > 0.1, which means that they contributed very little information to the test. All of the other items had maximum information scores ranging from 0.110-1.533, indicating a range of contribution to the evaluation of feigning. The item information functions were summed to create the test information function, which resulted in maximum information of 0.6 SD on the latent construct. This implies that the M-FAST performed best when the examinee had a latent trait level of 0.6 SD above the mean. An item that functioned at a similar latent trait level is Item 15 (b = .64 When I hear voices I hear them from either my right or my left ear, but rarely from both at the same time). As can be expected, the error increases and the amount of information decreases towards the extremes of the latent construct.

Rinaldo (2005) further calculated the relationship between the raw score on the M-FAST and the IRT score along the latent construct, which resulted in a correlation of .98 and showed near perfect match between these scores. Because item invariance is a fundamental aspect of IRT, invariance was calculated for the difficulty and slope. Rinaldo found invariance (i.e., the items measured the same underlying construct) for difficulty (r = .73) but not for slope (r = .04). Rinaldo opined that his smaller sample size may have contributed to the lack of invariance for the slope. In order to further evaluate invariance, Rinaldo calculated the parameter difficulty and
slopes again without the four problematic items. Again, there was invariance for difficulty ($r = .85$), but the slope decreased, indicating a poor fit than with the four problematic items ($r = .25$). After this further analysis, Rinaldo concluded that the low score on the slope correlation was not due to a small sample size but instead due to real noninvariance over malingering levels. He opined that the reason for this finding may be that examinees with mild levels of malingering may tend to endorse mood symptoms and examinees with higher levels of malingering may endorse more psychotic symptoms. Overall, Rinaldo stated that the M-FAST met IRT assumptions and had adequate IRT model fit.

**Purpose of the Present Study**

After reviewing literature on symptom feigning and sex differences, it is apparent that there are some gaps. One serious limitation is that, although some studies included males and females in the sample, with the exception of one study reported by Smith (2008), there was no analysis of sex differences. Smith (2008) found no significant differences between males and females on the total score of the M-FAST; however, there has been no study conducted to analyze sex differences on individual items.

Rinaldo (2005) conducted an Item Response Theory analysis on the M-FAST and found that, overall, the measure conformed to IRT assumptions and could be further evaluated using this analytic procedure. The reason the analysis did not result in a better fit was that four items (Items 2, 5, 11, and 23) had conflicting results. Three of these four items are questions based on mood symptoms. Because Rinaldo failed to analyze the data by sex, it is unknown whether the lack of conformity to IRT assumptions reflected differences between males and females.

The purpose of this study was to examine the effect of sex on the functioning of each item on the M-FAST. Furthermore, another aim of this study was to evaluate whether systematic
sex differences on individual items would lead to a biased increase in the number of females who are classified as potential malingerers based on the M-FAST cutoff score of 6. With the increasing number of females coming into contact with the legal system and the push for use of standardized, reliable assessments in malingering evaluations, it is important to understand how valid and reliable current malingering assessments are with this population.

**Hypothesis**

The main hypothesis for this study was that males and females would systematically score differently on the mood symptom items of the M-FAST. I expected that females would endorse these items more frequently than would males because of the higher rate of depression among females. Specifically, it was expected that women would have a lower latent trait score on mood Items 2, 5, and 23 compared to males. Mood Item 3 was not included in the analysis because endorsement of this item does not accurately reflect endorsement of a mood symptom due to the follow-up question and the lack of poor fit in the research for this item. This means that the difficulty \( (b) \) of the item, which is defined as when the probability of endorsing the item is 0.5, would be lower for females when compared to males.
Method

Participants

The participants were obtained from three different sources: two previous studies (Hill, 2009, and Montes, 2012), and a third sample used specifically for the purpose of this study for a total of 423 participants with 50 participants from Sample 1, 102 participants from Sample 2, and 271 participants from Sample 3. The final analysis utilized data from 407 participants.

Sample 1. The original dataset from Hill (2009) consisted of 50 male and 50 female correctional inmates on intake status at an Oregon prison; however, only half of the dataset was available for inclusion in the current study, resulting in a subsample of 25 males and 25 females. All of the participants were English-speaking. Data was not available concerning potential participants who refused to participate in the sample. A short demographic questionnaire, the M-FAST and the SIRS were administered to the participants. The M-FAST and the SIRS were counterbalanced during the administration. The PAI, which was administered by ODOC staff, was also utilized in this study. The participants were instructed to answer the items honestly.

Sample 2. The participants from Montes (2012) consisted of 102 incarcerated bilingual (English- and Spanish-speaking) Hispanic males. In this study, two bilingual potential participants refused to participate in the study (Montes & Guyton, 2014). A short demographic questionnaire, an acculturation measure, and the M-FAST in both English and Spanish were administered to the participants. The English and Spanish versions of the M-FAST were counterbalanced during administration. In this study, participants were randomly assigned to one of three conditions: two coached groups and one honest-responding group. These are the same experimental groups utilized in this study. The instructions given to participants in each of the three groups are as follows (p. 69-70).
**Honest condition.** Now, I am going to ask you some questions about your mental health. It is very important that you respond honestly to every question. Please answer every question.

**Coached condition.** Now, I am going to ask you some questions about your mental health. What I would like you to do is pretend and act as if you have a serious mental illness. Before I start asking you these questions, you can take a moment to think about how you will answer the questions to appear “crazy.” Please try to be as believable and convincing as possible. I really want you to convince me that you are “crazy.”

**Coached and warned condition.** Now, I am going to ask you some questions about your mental health. What I would like you to do is pretend and act as if you have a serious mental illness. Before I start asking you these questions, you can take a moment to think about how you will answer the questions to appear “crazy.” Please try to be as believable and convincing as possible, but be very careful and try not to overly exaggerate your mental health problems. I really want you to convince me that you are “crazy.”

These instructions were also given in Spanish. For the purposes of this study, only the results of the English administration of the M-FAST will be utilized from this sample.

**The current study.** The third sample specific to this study consisted of 271 inmates incarcerated at an Oregon prison. Of the 271 participants, eight failed to complete the M-FAST, six failed the manipulation check questions (i.e., failed to endorse using the instructions when answering the questions) and one identified as transgendered\(^1\). Data for these individuals were removed from the dataset, and data for the remaining 256 participants were used in the analysis.

---

\(^1\) The individual who identified as transgendered was not included in the final analysis because the analysis requires a dichotomous variable. Furthermore, it is likely that a transgendered individual may have a different approach to the research question and alter the statistics.
There were multiple potential participants who declined to participate. Sixty-five females refused to participate. The data for males who declined is unavailable. There being significantly more males on intake status than females (4 male intake units versus a half unit for females on intake status). Therefore, there was no record kept to ensure that he would not be approached again to participate in the study. For the 65 female potential participants who declined, the average age was 33, and the vast majority identified as White/Caucasian (77.3%).

Sample Characteristics

The total sample for this study was comprised of three separate samples combined into one. Table 1 displays the frequencies of the demographic variables for each sample individually and the total sample. An observational review of the demographic characteristics for Sample 1 suggests relatively comparable frequencies between the sexes on all variables. The sample tended to be under 40 years old, although more females identified as 50 and older. The majority identified as White/Caucasian, which is consistent with Oregon state\(^2\) as well as Oregon Department of Corrections census data. The majority of participants in this sample obtained a high school diploma/GED; however, females were slightly more likely to have a grade school/some high school education than were males. There were similar percentages for number of children between males and females, with the majority of both sexes reporting one to two children. The majority of males and females identified as single, never married; however, males were 24% more likely to endorse this option, and females tended to endorse the remaining options slightly more often than males.

\(^2\) According to the 2010 census, in the state of Oregon, 78.5% of the population identified as White alone, 11.7% identified as Hispanic or Latino, 1.7% identified as Black or African American alone, 1.1% as American Indian and Alaska Native alone, 3.6% as Asian alone, 0.3% as Native Hawaiian and Other Pacific Islander alone, 0.1% as Some Other Race alone, and 2.9% as Two or More races (Profile of General Population and Housing Characteristics: 2010 Demographic Profile Data, 2014).
As previously indicated, Sample 2 consisted of only males who, as expected, predominately identified as Hispanic. Over half of Sample 2 identified as 18-29 years old, approximately 20% more than Sample 1. Half of the sample endorsed receiving grade school/some high school education and almost half (46%) endorsed having a high school diploma/GED. Similar to Sample 1, participants in Sample 2 predominately reported having one to two children, although they were slightly more likely to report having no children as opposed to Sample 1. Approximately three fourths of the participants identified as single, never married which was 16% more than the males in Sample 1 and 38.5% more than the females in Sample 1.

The majority of the participants from Sample 3 identified as White/Caucasian and under 40 years old, similar to Sample 1. A majority of males reported having a high school diploma/GED, although approximately one fifth reported having grade school/some high school and one fifth reported having some college. For females, the percentages were more evenly distributed; approximately 30% reported having grade school/some high school, high school degree/GED, and some college. Approximately half of the males reported having one to two children, 30% having zero children and 20% having three to five children. In contrast females reported having one to two children and having three to five children at similar rates (39% and 35.6%). Approximately 20% of the females reported having no children. Roughly half of the males in this sample reported being single, never married; 30% reported being divorced and 11% indicated that they were currently married. Although more females reported being single, never married, the total percentage was lower than that for males at 40%. Roughly 20% of the female sample reported being married or divorced, and roughly 12% reported being separated.

The total sample yielded similar percentages to Sample 1 and Sample 3. The vast majority of both males and females were under the age of 40. Due to Sample 2 consisting solely
of Hispanic males, approximately half of the total male sample identified as Hispanic and roughly 30% identified as White/Caucasian as compared to approximately 70% of the total female sample who identified as White/Caucasian. A z-test was conducted to determine if the difference between the ethnicities was statistically significant based on sex. The results of the z-test indicated there was a statistically significant difference for ethnicity by sex (z = 523.27, p < 0.05). Slightly less than half of the males reported having a high school diploma/GED, and 35% reported having a grade school/home high school and 13% reported having some college. In comparison, females were more evenly distributed across the those three categories with 37% reporting having a high school diploma/GED, and roughly 30% reporting both grade school/some high school and some college. Roughly 40% of both males and females reported having one to two children. Approximately a quarter of the males reported having no children and a quarter reported having three to five. Females were slightly less likely to report having no children and slightly more likely to report having three to five children. Both sexes were more likely to endorse being single, never married; however, males endorsed this category roughly 20% more than the females. The percentage of males and females who endorsed being married and divorced were comparable amongst the sexes, females were slightly more likely to endorse both categories over males.
## Table 1

**Demographic Characteristics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample 1 (N=25)</th>
<th>Sample 2 (N=102)</th>
<th>Sample 3 (N=177)</th>
<th>Total (N=255)</th>
<th>Total (N=205)</th>
<th>Total (N=202)</th>
<th>Total (N=407)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male (N=25)</td>
<td>Female (N=25)</td>
<td>Total (N=50)</td>
<td>Male (N=78)</td>
<td>Female (N=177)</td>
<td>Total (N=255)</td>
<td>Male (N=25)</td>
</tr>
<tr>
<td></td>
<td>f %</td>
<td>f %</td>
<td>f %</td>
<td>f %</td>
<td>f %</td>
<td>f %</td>
<td>f %</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>8 32%</td>
<td>6 24%</td>
<td>14 28%</td>
<td>30 38.5%</td>
<td>65 36.7%</td>
<td>95 37.3%</td>
<td>93 45.4%</td>
</tr>
<tr>
<td>30-39</td>
<td>8 32%</td>
<td>8 32%</td>
<td>16 32%</td>
<td>21 26.9%</td>
<td>56 31.6%</td>
<td>77 30.2%</td>
<td>59 28.8%</td>
</tr>
<tr>
<td>40-49</td>
<td>7 28%</td>
<td>6 24%</td>
<td>13 26%</td>
<td>11 14.1%</td>
<td>46 26%</td>
<td>57 22.4%</td>
<td>33 16.1%</td>
</tr>
<tr>
<td>50-59</td>
<td>2 8%</td>
<td>5 20%</td>
<td>7 14%</td>
<td>13 16.7%</td>
<td>8 4.5%</td>
<td>21 8.2%</td>
<td>17 8.3%</td>
</tr>
<tr>
<td>60+</td>
<td>0 0%</td>
<td>0 0%</td>
<td>0 0%</td>
<td>0 0%</td>
<td>3 3.8%</td>
<td>2 1.1%</td>
<td>5 1.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>2 8%</td>
<td>0 0%</td>
<td>2 4%</td>
<td>3 3.8%</td>
<td>12 6.8%</td>
<td>15 5.9%</td>
<td>5 2.4%</td>
</tr>
<tr>
<td>Asian</td>
<td>0 0%</td>
<td>0 0%</td>
<td>0 0%</td>
<td>0 0%</td>
<td>1 1.3%</td>
<td>2 1.1%</td>
<td>3 1.2%</td>
</tr>
<tr>
<td>American Indian/</td>
<td>0 0%</td>
<td>1 4%</td>
<td>1 2%</td>
<td>4 5.1%</td>
<td>8 4.5%</td>
<td>12 4.7%</td>
<td>4 2%</td>
</tr>
<tr>
<td>Alaska Native</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>4 16%</td>
<td>2 8%</td>
<td>6 12%</td>
<td>100 98%</td>
<td>7 9%</td>
<td>10 5.6%</td>
<td>17 6.7%</td>
</tr>
<tr>
<td>Caucasian</td>
<td>19 76%</td>
<td>21 84%</td>
<td>40 80%</td>
<td>0 0%</td>
<td>53 67.9%</td>
<td>122 68.9%</td>
<td>175 68.6%</td>
</tr>
<tr>
<td>Bi/Multiracial</td>
<td>0 0%</td>
<td>1 4%</td>
<td>1 2%</td>
<td>2 2%</td>
<td>8 10.3%</td>
<td>22 12.4%</td>
<td>30 11.8%</td>
</tr>
<tr>
<td>Other</td>
<td>0 0%</td>
<td>0 0%</td>
<td>0 0%</td>
<td>0 0%</td>
<td>2 2.6%</td>
<td>1 0.6%</td>
<td>3 1.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Education Level</strong></td>
<td>3 12%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade/ Some High School</td>
<td>6 24%</td>
<td>9 36%</td>
<td>15 30%</td>
<td>51 50%</td>
<td>15 19.2%</td>
<td>50 28.2%</td>
<td>65 25.5%</td>
</tr>
<tr>
<td>Diploma/GED</td>
<td>13 52%</td>
<td>11 44%</td>
<td>24 48%</td>
<td>2 46%</td>
<td>37 47.4%</td>
<td>64 36.2%</td>
<td>101 39.6%</td>
</tr>
<tr>
<td>Some College</td>
<td>5 20%</td>
<td>4 16%</td>
<td>9 18%</td>
<td>5 4.9%</td>
<td>17 21.8%</td>
<td>53 29.9%</td>
<td>70 27.5%</td>
</tr>
<tr>
<td>College Deg.</td>
<td>1 4%</td>
<td>1 4%</td>
<td>2 4%</td>
<td>0 0%</td>
<td>9 11.5%</td>
<td>10 5.6%</td>
<td>19 7.5%</td>
</tr>
</tbody>
</table>

---

3 The data collection method for this item was inconsistent across the three samples as Sample 1 collected interval data (i.e. number of years spent in school) while Samples 2 and 3 collected nominal data (i.e. grade school, some high school; high school diploma/GED; some college; college degree). The data from Sample 1 was transformed into nominal data using the following criteria: < 12 years of school became grade school, some high school; 12 years became high school diploma/GED; 13 to 15 years of schooling became some college; and 16 + years became college degree.
## Number of Children

<table>
<thead>
<tr>
<th>Range</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>20%</td>
</tr>
<tr>
<td>1-2</td>
<td>12</td>
<td>48%</td>
</tr>
<tr>
<td>3-5</td>
<td>6</td>
<td>24%</td>
</tr>
<tr>
<td>6+</td>
<td>2</td>
<td>7%</td>
</tr>
</tbody>
</table>

## Legal Marital Status

<table>
<thead>
<tr>
<th>Status</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never Married</td>
<td>15</td>
<td>60%</td>
</tr>
<tr>
<td>Married</td>
<td>5</td>
<td>20%</td>
</tr>
<tr>
<td>Separated</td>
<td>1</td>
<td>4%</td>
</tr>
<tr>
<td>Divorced</td>
<td>4</td>
<td>16%</td>
</tr>
<tr>
<td>Widowed</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>
Measures

Demographic Questionnaire. A short demographic questionnaire was administered to participants in Sample 3 only. Different demographic questionnaires were administered in the other two Samples. This measure consists of information on age, sex, marital status, ethnicity, education level and previous mental health treatment. The entire questionnaire is available in Appendix A.

Miller Forensic Assessment of Symptoms Test (M-FAST). The M-FAST (Miller, 2001) is a 25-item structured interview designed to be used as a screening tool for malingering. It takes approximately 5 min to administer. Items are scored as “0” or “1.” Scores range from 0 to 25, with a cutoff score of 6 to suggest feigning (Miller, 2001). This test utilizes seven strategies for detection: unusual hallucinations, reported verses observed symptoms, extreme symptomatology, rare combinations, negative image, unusual symptom course, and suggestibility (Miller, 2001). This measure has acceptable psychometric properties, as reviewed earlier. The M-FAST contains four items with content related to mood state, Item 2, 3, 5, and 23. There was no rationale in literature on the development of the measure for why these questions were added (Miller, 1999; Miller, 2001).

Manipulation Check Questionnaire. The manipulation check questionnaire was comprised of four questions aimed at assessing the participant’s attention to and implementation of the instructions read prior to completion of the M-FAST. The participant was asked to read and answer the questions.

Procedure

All procedures for all three samples were approved by the Institutional Review Board of Pacific University and Oregon Department of Corrections. Participants for each of the three
samples were randomly selected from intake in the Oregon Department of Corrections. The experimenter in each study obtained a list of inmates on intake status from ODOC staff. The experimenters approached the potential participant and briefly explained the study. When the participant agreed to continue, he or she was escorted to an interview room. The procedure began with a conversation about informed consent and the experimenter answered any questions the participant had. The rest of the procedural explanation is specific to the sample for this study only because the other samples were required to complete additional measures specific to the research question for that study (refer to the individual studies for a more detailed account of the procedures that were utilized). As previously noted, the M-FAST was counterbalanced when administered in the other two samples. The participant was given the brief demographic questionnaire. Next, the participant was randomly assigned to one of the three coaching samples and the proper instructions were given to him or her. The instructions that were used were the instructions previously stated under Sample 2. The M-FAST was then administered. Finally, the participant was given the manipulation check and debriefed as to the purpose of the study and referred to any available services if needed such as individual sessions, or group therapy. No participants requested or were referred to services. Testing sessions lasted between 10 to 15 min.

Data Analysis

In order to fully understand Item Response Theory (IRT), and its application to the construct of symptom feigning, a brief history of the theory is necessary. IRT was developed in the 1940s; however, it has not been applied to psychological research until the last 10 or 15 years (Harvey & Hammer, 1999). This is due in part to the demanding and expensive computations that are required. Therefore, during its inception and prior to easily accessible computer software scoring, IRT was mostly used with large-scale standardized aptitude and achievement testing.
The specific use of aptitude and achievement testing has greatly influenced IRT vernacular, which sometimes leads to confusion when expanding IRT to other constructs, such as feigning. With this in mind, great care will be taken to explain the construct of feigning in the context of IRT.

IRT provides considerable benefits to the construction and analysis of psychological measurements when compared to Classical Test Theory (CTT) that will be addressed throughout this paper. IRT will be compared to CTT because the majority of psychological measures currently being utilized in the field have been created using CTT. The cornerstone of IRT is based on analysis of individual items as opposed to a total or raw scale score, as in CTT, because each item taps into the latent construct (denoted as the Greek letter theta, $\theta$). A probability can be calculated to determine where along that continuum the person falls. For example, if a person has low ability on the latent construct, then that person will have low probability of endorsing certain items. There are three main components in IRT: item characteristic curves (ICCs), item information functions, and invariance.

Backer (2001) stated that an ICC is a graph of the interaction between a person’s latent trait (X-axis) and probability of endorsing the item (Y-axis). The graph tends to be S-shaped to denote that at each ability level, there is a certain probability of endorsing that item which is known as the probability of theta, $P(\theta)$, as seen in Figure 1. The S-shape shows that people low on the latent trait will have a lower probability of endorsing the item and people high on the latent trait will have a higher probability of endorsing the item. Every item in the test will generate an ICC.
Baker (2001) described the interpretation of the ICC as being composed of two functions of the graph: difficulty and discrimination parameters. Although difficulty and discrimination are used in CTT to describe items, in IRT, definitions of these parameters are based in theory and carry different meanings. In IRT the difficulty of an item represents where that item falls along the latent trait continuum; in other words, whether the person falls on the high end or the low end. In the context of aptitude and achievement tests, this term makes sense; however, this term may be confusing when discussing the latent trait of feigning because feigning is not a trait but a state. For the construct of feigning, the descriptors of low and high are not applicable. Therefore, it is best to conceptualize people low on the latent trait of feigning as people who are exaggerating only a few symptoms and to conceptualize people high on the latent trait of feigning as exaggerating many symptoms. If an item has low difficulty, people low on the latent

---

4 All graphs in the Data Analysis section were made with factitious data using R software (Retrieved from http://www.r-project.org).
trait will have a low probability of endorsing the item. Consequently, people who are exaggerating only a few symptoms would have a lower probability of endorsing an item with low difficulty.

Similarly, if an item has high difficulty, people exaggerating many symptoms will have a high probability of endorsing the item. Difficulty is denoted by the Greek letter beta ($b$) and is the point on the latent trait continuum where the $P(\theta) = 0.5$. This means that at the given latent trait level, the probability of endorsing the item is 50%. In theory, the value of the difficulty parameter ranges from negative infinity to positive infinity and is asymptotic at a probability of 0 and 1. In practice, the value ranges from $-3 < b < +3$ (Baker, 2001). The item difficulty parameter is analogous to the item mean in CTT (Reise, Ainsworth, & Haviland, 2005). Figure 2 shows four ICCs next to each other with different $b$ values. The ICC on the far left depicts an item that has low difficulty because at $P(\theta) = 0.5, \theta = -1$. In contrast, the ICC on the far right depicts an item that has high difficulty because at $P(\theta) = 0.5, \theta = 1.0$.

Figure 2: ICCs for Item Difficulty
Baker (2001) described discrimination as how well the item differentiates between people who score below and above the item location (aka the difficulty). Discrimination can also be thought of as the steepness or the slope of the ICC. If the slope is steep, then the item does a good job discriminating between people low and high on the latent trait, and, if the slope is flat, then the item discriminates poorly. If the discrimination is poor that means that the item has high probability of endorsement by both low and high scorers, which means it does not show a difference between high scorers and low scorers. The other way discrimination can be poor is if the item is rarely endorsed by either high or low scorers. Discrimination is denoted by the Greek letter alpha ($\alpha$) and represents the area along the S-shape where $\theta = b$ and where the slope is the steepest. Once again, in theory the value of this parameter also ranges from negative infinity to positive infinity, but in practice it tends to range from $-2.8 < \alpha < +2.8$ (Baker, 2001). The item discrimination is analogous to item-test correlation in CTT or factor loading in factor analysis (Reise et al., 2005). Figure 3 shows four ICCs with varying $\alpha$ values with $b$ held constant. The ICC with low discrimination (Item 4) has mostly lost the characteristic S-shape of the ICC. The ICC with high discrimination (Item 1) has a sharp S-shape with a rapid change between the people low and high on $\theta$. 

The second fundamental part of IRT is item information functions (Baker, 2001). Baker (2001) reported that the item information function is used to judge the quality of the item. Different items relay varying amounts of information about different ranges on the latent continuum. There is also an item function for each response on an item. For example, items with low difficulty (items on which individuals who have low levels of feigning have a lower probability of endorsing) relay information about how individuals low on the latent trait score along the latent continuum, and items with high difficulty (items on which individuals who have a high level of malingering have a higher probability of endorsing) relay information about how individuals high on the latent trait score. Items with low differentiation provide less information than items with high differentiation. Maximum information is obtained around the difficulty value. The item information functions can then be summed to equal the scale information function, which is a direct measure of the precision of the scale to measure the latent construct. In order to conceptualize this part of IRT it may be helpful to think of the item information
function as analogous to item reliability in CTT and the scale information function is analogous to scale reliability in CTT (Reise et al., 2005). Figure 4 shows four Information Function Curves (IFC). Item 1 contributes little information and has a low, wide bell curve. Item 4 contributes the most information out of the four items and has a taller, narrow bell curve.

Figure 4: Information Function Curves

![Information Function Curves](image)

The last main component of IRT is invariance. Reise et al. (2005) described invariance as a fixed score along the latent construct. The latent construct is the underlying trait that is being measured. For example, questions about changes in sleeping patterns and appetite are ways to measure a person’s depression. In this example, depression is the underlying trait or latent construct. The idea of invariance is addressed throughout IRT and therefore, will be addressed in the context of the previously stated parts of IRT. Item difficulty is one of the main ways of describing the ICC. With CTT the difficulty of an item is determined by the group taking the test. Therefore, for people high on the latent trait, the item would not be difficult, but for people low on the latent trait the item would be difficult. There would be no way of knowing the true
difficulty of the item. In IRT, the difficulty of the item is where that item falls along the latent continuum and therefore is not affected by different group performance. This is true for discrimination as well. Under CTT item discrimination is based upon the group, and in IRT it is based upon the latent continuum.

Reise et al. (2005) stated that the idea of invariance can be taken a step further to say that the examinee’s latent trait is invariant of the items on the test. The best way to understand this concept is with an example. An examinee takes a test that has five easy items, which computes a score along the latent trait of $\hat{\theta}_1$. The examinee takes another test, which measures the same latent construct but has five hard items which computes a score along the latent trait of $\hat{\theta}_2$. Because the latent construct has invariance and the ICC spans the entire range of the latent trait continuum, $\hat{\theta}_1 = \hat{\theta}_2$. If this same scenario was carried out with CTT, the examinee would obtain a high score on the test with easy items and a low score on the test with hard items and there would be no way of knowing where the examinee falls along the latent trait continuum.

According to Reise et al. (2005), one of the biggest advantages of IRT is based upon the item parameters being independent and the examinee’s latent trait being independent. An individual’s score on the measure is not a sum of the items of the measure but instead where the individual falls along the latent construct. This also influences item discrimination. In CTT, item discrimination is based on how many individuals get the item correct. In IRT, discrimination is the way an item performs over different levels of the latent construct. Invariance allows for comparisons among different groups on the same latent construct, even across various measures. When using tests developed with CTT, it is difficult to determine whether different groups such as different age groups, or ethnic groups systematically score differently on the measure. With IRT, groups can be compared based on where they fall along the latent trait, even if the groups
systematically differ on the test. Even if different groups experience different symptoms of the latent trait or interpret items of a test differently, their estimated latent trait score will allow researchers to compare apples to apples which is not an option under CTT.

Differential item analysis (DIF) is the procedure used to determine whether groups score differently on the test items (Osterline & Everson, 2000). According to Osterline and Everson (2000), if it is determined that one group scores significantly different than another group, then bias needs to be evaluated. It may be that one group has more of the latent construct by nature or the groups are not expected to score similarly; then the difference is not a bias. However, if the “examinee responses to particular test items or sets of test items are linked systematically to the personal characteristics (such as sex or ethnicity) of the examinees and are otherwise unrelated to the test’s central construct” (p. 3), then the test is biased against a certain group. If the assessment is being created, then the bias can be eradicated before the assessment is finalized. If the assessment is already in use, then the author(s) can leave the assessment as is or create a new edition that address the problem.

IRT models. In order to create the ICC, the parameters need to be calculated to determine the best model fit for the data. According to Baker (2001), there are three IRT models: the one-parameter logistic model or the Rasch Model, the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL). All of the models assume that the latent construct is a determinant of the examinee’s response to an item (Harvey & Hammer, 1999). The models are different in how the latent construct causes the item response which is defined by the item parameters (Harvey & Hammer, 1999). The models also take into account dichotomous versus graded responses.
According to Baker (2001), the Rasch model predominately utilizes the difficulty \((b)\) parameter to describe the item. Although \(\alpha\) is still a part of the ICC, it remains constant across all of the items at \(\alpha = 1.0\) and \(b\) varies among the items. This means that the shape of ICC will stay the same; the only variable that would change would be how far right or left the line falls.

Baker (2001) reported that the 2PL model looks at the two parameters of difficulty and discrimination. It takes the Rasch model and allows \(\alpha\) to vary. In this model, the ICC can move from right to left, and the slope of the line can change. The slope reflects the degree of relationship between the item and the latent construct. For example, items with larger \(\alpha\) values have a stronger relationship to the latent construct, whereas items with smaller values have a weaker relationship (Harvey & Hammer, 1999).

Baker (2001) described the 3PL model has the same parameters as the 2PL, but it adds a third guessing parameter labeled \(c\). Dichotomous tests that utilize multiple-choice, true-or-false, or yes-or-no questions leave the possibility open that an examinee can correctly guess the answer. The \(c\) parameter raises the lower asymptote of the ICC. By adding \(c\), the value and definition of \(b\) also changes. Under the 1PL and 2PL models, the value of \(b\) was defined as the value of \(\theta\) that had a probability of .50. Now that the lower asymptote is changed, the new definition of \(b\) is the probability that lies halfway between \(c\) and 1. By increasing the lower asymptote, the discrimination of the item \((a)\) and the difficulty of the item \((b)\) are changed. As \(c\) increases, the item becomes less discriminative and more difficult and therefore reduces the amount of information presented by the item. Put another way, as an item becomes easier to guess, the less information the item yields for estimating an examinee’s score on the latent construct. The more difficult it is for an examinee to guess the right answer, the more information a right or wrong score yields.
Data analysis for current study. The M-FAST scores were analyzed using the Differential Item Functioning model with IRTPRO software from Scientific Software International. Items were analyzed for their ICCs, and IFCs. Next, a DIF analysis was conducted on all items to determine invariance between males and females. A statistically significant difference between the groups is determined by the difference between the log likelihood of the item parameters for the data analyzed as a single group and the log likelihood of the item parameters for the separate groups (du Toit, 2003). The difference and degrees of freedom are then located in a chi-square distribution.

The manipulation check was analyzed using a chi-square test to ensure that participants understood the instructions and then used that strategy when approaching the test. Participants who failed the manipulation check were removed from the data set.
Results

M-FAST Item Endorsement Trends

Prior to completing the IRT analysis of the M-FAST items, trends in item endorsement for the total sample were reviewed. Table 2 displays the frequencies of item endorsement by sex for each category. The data was interpreted through observation. Males and females in the honest condition endorsed items in a similar manner, except for Items 21 and 23, on which females were more inclined to endorse the items than males. In the coached condition, males were more likely to endorse items overall than females. For the coached and warned category, females were more likely overall to endorse items than males, except for items 5 and 15 in which males more likely to endorse the item. For Items 4, 6, 7, 24, and 25, females endorsed the items 20-35% more than the males in this category. Endorsement rates across the items for all males and all females indicate relatively consistent endorsement, except for Item 12 in which females endorsed the item 10% more often.
Table 2

**Frequency of Endorsement of M-FAST Items per Category by Sex**

<table>
<thead>
<tr>
<th>Items</th>
<th>Honest (N=67)</th>
<th>Coached (N=69)</th>
<th>Coached and Warned (N=67)</th>
<th>Total (N=202)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>23.9%</td>
<td>27</td>
<td>40.3%</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>14.9%</td>
<td>15</td>
<td>22.4%</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3%</td>
<td>3</td>
<td>4.5%</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>26.9%</td>
<td>17</td>
<td>25.4%</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0%</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.5%</td>
<td>3</td>
<td>4.5%</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1.5%</td>
<td>3</td>
<td>4.5%</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1.5%</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1.5%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0%</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>6%</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1.5%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>6%</td>
<td>5</td>
<td>7.5%</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1.5%</td>
<td>4</td>
<td>6%</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>3%</td>
<td>3</td>
<td>4.5%</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>6%</td>
<td>5</td>
<td>7.5%</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0%</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1.5%</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1.5%</td>
<td>5</td>
<td>7.5%</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>3%</td>
<td>10</td>
<td>14.9%</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>1.5%</td>
<td>4</td>
<td>6%</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>11.9%</td>
<td>15</td>
<td>22.4%</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>0%</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>0%</td>
<td>2</td>
<td>3%</td>
</tr>
</tbody>
</table>

*Note. Items of interest are shown in boldface*
The specific mood items were reviewed for trends in frequency of endorsement across sex. For Item 2, females were slightly more likely (22.4%) to endorse this item in the honest group and even more likely to endorse the item in the warned group (82.1%) compared to males (14.9% and 66.7%). Males and females were fairly consistent in the endorsement rates in the coached category, with males endorsing the item 3.2% more often than females. For the total sample, females were 6.3% more likely to endorse the item across the categories than males. For Item 5, males and females endorsed the item at similar rates in the honest (1.5%) and coached categories (5.1%); however, males were 15% more likely to endorse the item in the warned category than females. In the total sample, males endorsed the item 7.3% more often than did females. For Item 23, females endorsed the item 10.5% more often in the honest group than did males, and 25.3% more often in the warned group than did males. Males and females demonstrated similar rates of endorsement in the coached category. In the total sample, females endorsed the item 8.7% more overall than males.

For this study, the participants were provided instructions on how to approach the measure to ensure that all levels of the latent trait were represented. The M-FAST total scores are presented in Table 3. The majority of the honest responders for both males and females scored below the cut off score of 6 as expected for honest responders. For the coached and coached and warned responders, the majority of both males and females scored above the cut off score of 6, as expected and represented both moderate and extreme ranges of feigning.
among the assumption of unidimensionality is that there is only one latent trait to explain the variance

IRT Analysis

Assumption of unidimensionality

Dimensionality refers to the number of latent traits that are represented in the data. The assumption of unidimensionality is that there is only one latent trait to explain the variance among the item responses (Embretson & Reise, 2000). Unidimensionality was determined for
this data set using the $M_2$ test which was significant ($M_2(275) = 498.64, p < .001$), suggesting that the 25 M-FAST items were all from the same scale.

**DIF analysis**

The analysis began with an unconstrained baseline model in which the mean level of the underlying trait was allowed to vary across the two sexes (i.e., a standard 2PL model). In order to compare the items, the item parameter estimates needed to be calibrated to the same scale. This was done by identifying “anchored” items and using these items to estimate the parameters. Anchored items are items that are determined to have no bias. Based on the information from Rinaldo (2005) the following items from the M-FAST were used as anchor items for this analysis: 1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, and 25. 2PL models were run for both groups. Table 3 presents the results for each item. The table shows that the anchored items were not free to vary by group. The primary parameter of estimate of the models (difficulty, $b$) was similar for both groups. The largest difference in difficulty for the items was on Item 5, which had a 0.58 difference in difficulty between males and females.

ICC curves were created for each item for both groups. Figures 5 and 7 show that the difficulty for these items was moderate (around 0.00 for both groups) and that the discrimination for both of these items was high (2.15 to 2.5; Baker, 2001). However, as Figure 6 shows, the difficulty for Item 5 was high ($\geq 1.00$ for both groups), and the discrimination was very low. This suggests that Item 5 did not discriminate between the participants who were feigning and those who were not feigning. Thus, participants who scored highly overall only had a slightly higher probability of endorsing Item 5 than did those who scored low overall.
Table 4

*Results for 2PL Models with Anchored Items per Sex*

<table>
<thead>
<tr>
<th>Item</th>
<th>Males</th>
<th></th>
<th></th>
<th>Females</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$c$</td>
<td>$b$</td>
<td>$\alpha$</td>
<td>$c$</td>
<td>$b$</td>
</tr>
<tr>
<td>1</td>
<td>1.94</td>
<td>1.16</td>
<td>-0.60</td>
<td>1.94</td>
<td>1.16</td>
<td>-0.60</td>
</tr>
<tr>
<td>2</td>
<td><strong>2.65</strong></td>
<td><strong>0.36</strong></td>
<td><strong>-0.13</strong></td>
<td><strong>2.54</strong></td>
<td><strong>0.64</strong></td>
<td><strong>-0.25</strong></td>
</tr>
<tr>
<td>3</td>
<td>3.65</td>
<td>-0.25</td>
<td>0.07</td>
<td>3.65</td>
<td>-0.25</td>
<td>0.07</td>
</tr>
<tr>
<td>4</td>
<td>2.96</td>
<td>-2.25</td>
<td>0.76</td>
<td>2.96</td>
<td>-2.25</td>
<td>0.76</td>
</tr>
<tr>
<td>5</td>
<td><strong>0.33</strong></td>
<td><strong>-0.33</strong></td>
<td><strong>1.00</strong></td>
<td><strong>0.43</strong></td>
<td><strong>-0.68</strong></td>
<td><strong>1.58</strong></td>
</tr>
<tr>
<td>6</td>
<td>4.76</td>
<td>-1.10</td>
<td>0.23</td>
<td>4.76</td>
<td>-1.10</td>
<td>0.23</td>
</tr>
<tr>
<td>7</td>
<td>4.13</td>
<td>-1.18</td>
<td>0.29</td>
<td>4.13</td>
<td>-1.18</td>
<td>0.29</td>
</tr>
<tr>
<td>8</td>
<td>4.35</td>
<td>-0.66</td>
<td>0.15</td>
<td>4.35</td>
<td>-0.66</td>
<td>0.15</td>
</tr>
<tr>
<td>9</td>
<td>3.09</td>
<td>-1.80</td>
<td>0.58</td>
<td>3.09</td>
<td>-1.80</td>
<td>0.58</td>
</tr>
<tr>
<td>10</td>
<td>3.11</td>
<td>-1.40</td>
<td>0.45</td>
<td>3.11</td>
<td>-1.40</td>
<td>0.45</td>
</tr>
<tr>
<td>11</td>
<td>2.99</td>
<td>-1.21</td>
<td>0.41</td>
<td>2.99</td>
<td>-1.21</td>
<td>0.41</td>
</tr>
<tr>
<td>12</td>
<td>2.01</td>
<td>-1.11</td>
<td>0.55</td>
<td>2.01</td>
<td>-1.11</td>
<td>0.55</td>
</tr>
<tr>
<td>13</td>
<td>4.40</td>
<td>-0.61</td>
<td>0.14</td>
<td>4.40</td>
<td>-0.61</td>
<td>0.14</td>
</tr>
<tr>
<td>14</td>
<td>4.81</td>
<td>0.38</td>
<td>-0.08</td>
<td>4.81</td>
<td>0.38</td>
<td>-0.08</td>
</tr>
<tr>
<td>15</td>
<td>2.48</td>
<td>-0.66</td>
<td>0.27</td>
<td>2.48</td>
<td>-0.66</td>
<td>0.27</td>
</tr>
<tr>
<td>16</td>
<td>3.12</td>
<td>-0.47</td>
<td>0.15</td>
<td>3.12</td>
<td>-0.47</td>
<td>0.15</td>
</tr>
<tr>
<td>17</td>
<td>2.93</td>
<td>-0.47</td>
<td>0.16</td>
<td>2.93</td>
<td>-0.47</td>
<td>0.16</td>
</tr>
<tr>
<td>18</td>
<td>3.85</td>
<td>-0.58</td>
<td>0.15</td>
<td>3.85</td>
<td>-0.58</td>
<td>0.15</td>
</tr>
<tr>
<td>19</td>
<td>3.44</td>
<td>-0.67</td>
<td>0.19</td>
<td>3.44</td>
<td>-0.67</td>
<td>0.19</td>
</tr>
<tr>
<td>20</td>
<td>4.05</td>
<td>-0.39</td>
<td>0.10</td>
<td>4.05</td>
<td>-0.39</td>
<td>0.10</td>
</tr>
<tr>
<td>21</td>
<td>3.37</td>
<td>0.28</td>
<td>-0.08</td>
<td>3.37</td>
<td>0.28</td>
<td>-0.08</td>
</tr>
<tr>
<td>22</td>
<td>5.16</td>
<td>-0.21</td>
<td>0.04</td>
<td>5.16</td>
<td>-0.21</td>
<td>0.04</td>
</tr>
<tr>
<td>23</td>
<td><strong>2.36</strong></td>
<td><strong>-0.04</strong></td>
<td><strong>0.02</strong></td>
<td><strong>2.17</strong></td>
<td><strong>0.39</strong></td>
<td><strong>-0.18</strong></td>
</tr>
<tr>
<td>24</td>
<td>3.99</td>
<td>-1.63</td>
<td>0.41</td>
<td>3.99</td>
<td>-1.63</td>
<td>0.41</td>
</tr>
<tr>
<td>25</td>
<td>3.87</td>
<td>-1.10</td>
<td>0.28</td>
<td>3.87</td>
<td>-1.10</td>
<td>0.28</td>
</tr>
</tbody>
</table>

*Note.* $\alpha =$ discrimination, $b =$ difficulty, $c =$ guessing. Items that were not anchored are shown in boldface.
Figure 5. ICC for Item 2 for males and females.
Figure 6. ICC for Item 5 for males and females.
IFCs were also created for Items 2, 5, and 23. The curves for Items 2 and 23 showed moderate information was obtained for each of these items, with both having their peaks at a moderate latent construct score. However, the IFC for Item 5 showed that roughly no information was obtained from this item. The curves for both groups are presented in Figures 8 – 10.
Figure 8. IFC for Item 2 for males and females.
Figure 9. IFC for Item 5 for males and females.
Differential item functioning statistics were conducted to test for bias in the three items for the total sample of males versus females. Results of the DIF statistics showed a lack of significance for all three items ($p > 0.05$ for all). This suggests that there were no differences in how males and females responded to Items 2, 5, and 23. Thus, there was no bias between the sexes for the mood items. Results of the DIF statistics for the three items are presented in Table 4.
Table 5

Results for DIF Statistics for Items 2, 5, and 23 for the total sample

<table>
<thead>
<tr>
<th>Item</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.70</td>
<td>2</td>
<td>.689</td>
</tr>
<tr>
<td>5</td>
<td>2.70</td>
<td>2</td>
<td>.255</td>
</tr>
<tr>
<td>23</td>
<td>2.00</td>
<td>2</td>
<td>.364</td>
</tr>
</tbody>
</table>

Chi-Square Test

A post-hoc chi-square test was conducted on Items 2, 5, and 23 for males and females in the honest condition only to ensure that asking individuals to “pretend and act like you have a serious mental illness” did not alter how individuals genuinely approached the measure and therefore distorted any bias that may be present. There was no statistically significant difference between males and females on items 2, 5, or 23.

Table 6

Chi-Square Test Results for Items 2, 5, and 23 for the total honest condition

<table>
<thead>
<tr>
<th>Item</th>
<th>Pearson chi-square</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.270</td>
<td>.604</td>
</tr>
<tr>
<td>5</td>
<td>.362</td>
<td>.547</td>
</tr>
<tr>
<td>23</td>
<td>.027</td>
<td>.869</td>
</tr>
</tbody>
</table>
Discussion

Review and Implications of the Findings

The results of the DIF analysis showed that there was no statistically significant difference between males and females on any of the three mood items of the M-FAST included in this analysis. A post-hoc chi-square test for the participants in the honest responding group only corroborated this finding. This suggests that there was no sex difference on the mood items assessed. This finding was surprising given the research on gender differences for mood symptoms and the identification of poor fit for the mood items on the M-FAST.

One possible explanation for the lack of statistically significant difference between males and females on the mood items is the concerns that participants could have had when approached to participate in the study, regardless of sex, that may have impacted their answers to the M-FAST items. Morgan, Steffan, Shaw and Wilson (2007) identified self preservation concerns, which consisted of concerns regarding confidentiality and perceptions of weakness or colluding with staff, as barriers to receiving mental health treatment for incarcerated individuals. All of the participants in this study were interviewed when they were on intake status. It is possible that the participants were guarded when answering the questions posed to them due to concerns about confidentiality and how they may be viewed by the other inmates. Although data were not specifically collected on this topic for this study, researchers noted multiple behavioral observations regarding concern about confidentiality and being labeled a “mental health client” due to participation in the study. Furthermore, asking participants to “fake having a mental illness” could have increased these concerns due to the unconventional nature of the request.

The construct of symptom feigning and malingering is unique compared to the constructs of severe mental health disorders such as Major Depressive Disorder, Schizophrenia and Bipolar
Disorder. These disorders are diagnosed based on a set of symptoms that the individual is displaying. Malingering, on the other hand, is identified based on strategies an individual utilizes in certain situations with the two main strategies being unlikely presentation and amplified presentation (Rogers, 2008a). Based on the detection strategies, the content of the item (i.e., mood symptom versus psychotic symptom) is almost irrelevant. What is important is whether the symptom is a genuine symptom of psychopathology, is a rare or unusual symptom, if the individual is reporting the symptom to happen more often or more intensely than individuals with genuine psychopathology do, or if the individual is endorsing many items regardless of the content. When evaluating the construct of malingering, is it even reasonable to think that items with mood symptom content are consistent with genuine mood symptoms? The fact that research has shown that the M-FAST contains one main construct of “response styles indicative of malingering” (Miller, 2001, p. 30; Vitaco et al., 2008) would suggest no. If the mood symptom items were consistent with both constructs of depression and malingering, the measure would have multiple constructs. If the mood symptom items are not reflective of the depressive disorder construct, then it is not surprising that males and females did not approach the items the same way they would for genuine depressive disorders.

**Contribution of Items 2, 5, and 23 to the M-FAST**

The ICCs for Items 2 and 23 indicated that both of these items had moderate difficulty and high discrimination. The moderate difficulty of these items suggests that individuals who are not exaggerating, or exaggerating only a little are less likely to endorse the items, and individuals who are greatly exaggerating are more likely to endorse these items. The high discrimination of these items suggests that there is a clear and distinct difference between how individuals who are not exaggerating and those who are exaggerating approach the items. The combination of
moderate difficulty and high discrimination suggest that these items are helpful in identifying individuals who are feigning, which is corroborated by the IFCs for these items.

These parameters are inconsistent with the findings of Rindaldo (2005), who identified Items 2 and 23 as having poor fit. It is unclear at this point why these items performed so differently in the two separate studies. The construct of feigning/malingering is different than the construct of most diagnoses in that feigning/malingering is dynamic and can change based on the situation. Therefore, the different samples that were utilized may have greatly affected the items. For example, a small percentage of the sample for this study had some college education, approximately 25%, whereas all 600 participants in Rinaldo’s study were undergraduates. It is likely that these separate samples have a different understanding of mental illness. Furthermore, the M-FAST was designed to be used with populations that utilize malingering and malingering is more prevalent in a forensic setting than in a college counseling center. It is possible that the results of this study on Items 2 and 23 being a good fit was due to the sample consisting of forensic individuals in a forensic environment which is the population this measure was designed to be used with.

Item 5 displayed different results than Items 2 and 23. Although the difficulty of Item 5 was high, suggesting that individuals who are greatly exaggerating are more likely to endorse the item, the discrimination was extremely low, suggesting that it does not actually differentiate between individuals who are exaggerating and those who are not in a meaningful way. This is reflected in the IFC for Item 5, which suggests that the item contributes no helpful information related to the latent trait.

Item 5 has been consistently identified as a problematic item (Rinaldo, 2005; Vitacco et al., 2008). The results of this study suggest that the poor fit of this item is not a result of sex
difference. It is possible that this item fails to capture the latent trait. As previously stated, one of
the strategies for identifying feigned symptoms is by identifying exaggerated intensity or severity
of a symptom. Item 5 loads on the extreme symptomology scale. However, it is possible that the
wording of the item *I feel unusually happy most of the time* fails to tap into that construct, as it is
rather similar to the genuine symptom of mania, or that individuals are not responding to the
appropriate portion of the question. Furthermore, individuals who identify as being unusually
happy are typically not in distress. Because symptom feigning is typically the expression of
exaggerated distress, an item that fails to capture this aspect would likely contribute little to the
latent trait.

**Study Strengths and Limitations**

As with all studies, some strengths and limitations may have impacted the results and
conclusions. The sample utilized for this study was a combination of three separate samples, one
of which was comprised of bilingual Spanish speaking males. Montes and Guyton (2014)
evaluated the level of acculturation of the participants to assist with determining if the
participants who identified as Hispanic approached the measure in systematically different way
than say a Caucasian sample. His research indicated that the sample was highly acculturated and
approached the measure in a similar manner compared to the norming sample of the M-FAST.

Another potential limitation was the total sample size. For IRT analyses there is no easy
way to determine how many participants are necessary for the analysis as the necessary number
is based on a plethora of factors (Morizot, Ainsworth, & Reise, 2007). For a 2PL model, the
acceptable sample size has ranged between 200 and 500 participants (Morizot et al., 2007).
However, historically sample sizes for IRT studies have been in the 1,000s (Harvey & Hammer,
While the sample size for this study is deemed acceptable, it is possible that a larger sample size could reveal subtle nuances that were not detected with the current sample size.

Finally, there is the possibility that some of the participants responded negatively to the implications of “faking crazy” and/or being identified as a mental health inmate. This may have resulted in participants responding to the items in a more guarded manner and thus reducing the number and/or type of items they endorsed. Specifically, females may have been less likely to endorse the mood items out of fear or concern that their true symptoms may be identified and instead endorsed the more bizarre or odd items as a “safer” choice.

In the future, this issue could be reduced by adding more emphasis in the consent form specifically about confidentiality of the participant’s answers from DOC staff. Furthermore, in the year prior to collecting data for this study, I was a practicum student at the same facility. Although the participants of this study were all on intake status at the time, due to recidivism and access to general population inmates, it is possible that participants and potential participants learned of my previous role in the institution and impacted their self-presentation strategies. A way to reduce this confusion in the future would be to make sure that the researchers have no prior roles within the facility that may impact participant and potential participant’s view. Another way to reduce this issue could be to use general population inmates who may be more familiar with the procedures of the facility and possibly less suspicious of the researcher’s intentions.

Despite these limitations, this study is on the forefront of the growing demand for empirical research on how measures function for diverse groups. At the time of this writing, there are no IRT DIF evaluations for malingering measures published. This study is the first to evaluate for sex bias on the M-FAST. The results of this study demonstrate efficacy for using
this measure with both male and female offenders even though the measure was predominantly normed on males. This study also utilized a population in which the M-FAST is frequently utilized, thus increasing the generalizability of this study to forensic populations. Although the ethnic diversity of the sample was mentioned as a possible limitation of the study, it can also serve as a strength; the majority of incarcerated individuals are ethnic minorities.

**Recommendations for Future Research**

Based on the findings of this study and the limitations listed above, some recommendations for future research are proposed. As previously stated, there is limited research available on how different groups, and females in particular, perform on malingering or symptom feigning measures. Therefore, more research in general in this area would help to promote accuracy and specificity to important forensic referral questions as the majority of the individuals these measures are administered to are different than the norming populations.

The first recommendation would be to replicate the study with a substantially larger sample size. It would also be beneficial if the study was replicated in various geographic areas to capture a more diverse sample. The ODOC has less of an ethnically diverse population than the majority of the rest of the country. Therefore, replication of this study in different geographic areas could further increase the generalizability of the results. Replication of this study could also identify why Items 2 and 23 performed differently here than in Rinaldo’s (2005) study. It is possible that these items functioned differently in the two samples due to the difference in the sample characteristics. Therefore, IRT analyses of the M-FAST with different populations could assist in identifying why these items performed differently.

Another avenue of research would be to conduct DIF analyses on the M-FAST to determine if bias exists whether the measure for other diverse groups such as ethnicity or age. As
previously stated, the sample for this study consisted of predominantly Caucasian females and would have consisted of predominantly Caucasian males if Sample 2 was not included which consisted of Hispanic males. The majority of incarcerated individuals are ethnic minorities. It is likely that different ethnic minority individuals have different values and beliefs related to utilizing malingering and mental health in general.

On a larger scale, it is recommended that DIF analyses be conducted on other widely used malingering measures, especially the SIRS-2 given that it is considered the gold standard. Because the M-FAST was developed as a screening measure, a follow-up malingering evaluation is warranted if an individual scores above the cut-off score on the M-FAST. The diagnosis of malingering should not be made until a more thorough evaluation is completed. It is just as important, if not more important, to understand how different groups score on these measures as well.
Conclusions

The results of the DIF analysis suggests that the three items of focus for this study (Items 2, 5, and 23) of the M-FAST are free of sex bias; therefore, the measure performs equally well regardless of sex. Although this is an unexpected finding, it bodes well for the continued use of the M-FAST in forensic settings with both males and females. Due to the potential consequences and the stigma attached to the label of malingering or symptom feigning, it is all the more important for the measures used to identify these labels be of strong psychometric properties and perform equally well across a variety of individuals. Although the specific characteristics that define the forensic population are constantly in flux, females will continue to be represented in this population and are as likely to utilize feigning or malingering as males. Even though more research is needed in the area, the results of this study suggest that the M-FAST is a valid measure for screening this response style in females.
References


Appendix A

Demographic Questionnaire

Participant Number: ___________________

Date of Birth:___________________________

Age:_________________________________

Sex
   _____ Male
   _____ Female
   _____ Transgendered/Other

Race/Ethnicity
   _____ Caucasian/White
   _____ African-American
   _____ Hispanic/Latino/a
   _____ Asian-American
   _____ Native-American
   _____ Bi-/Multi-racial
   _____ Other: _______________________________

Highest level of education completed:
   _____ Grade school; last grade completed ______
   _____ High school diploma/GED
   _____ Some college; number of years completed ____
   _____ College degree; degree earned ___________

Legal Marital Status:
   _____ Single, never married
   _____ Married
   _____ Separated
   _____ Divorced
   _____ Widowed
   _____ Other: _______________________________

Number of times legally married:____________

Number of biological children:______________

Number of children living in your home before you were incarcerated: ________________

Number of incarcerations: __________________
Appendix B

Informed Consent

INSTITUTIONAL REVIEW BOARD
FWA: 00007392 | IRB: 0004173
2043 College Way | UC Box A-133 | Forest Grove, OR 97116

1. Study title
   Gender Differences on a Correctional Measure (196-12)

2. Study personnel

<table>
<thead>
<tr>
<th>Name</th>
<th>Megan Thomet</th>
<th>Michelle Guyton</th>
<th>Eloise Holdship</th>
<th>Jonathan Ryan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role</td>
<td>Principal Investigator</td>
<td>Faculty Advisor</td>
<td>Research Assistant</td>
<td>Research Assistant</td>
</tr>
<tr>
<td>Institution</td>
<td>Pacific University</td>
<td>Pacific University</td>
<td>Pacific University</td>
<td>Pacific University</td>
</tr>
<tr>
<td>Program</td>
<td>School of Professional Psychology</td>
<td>School of Professional Psychology</td>
<td>School of Professional Psychology</td>
<td>School of Professional Psychology</td>
</tr>
<tr>
<td>Email</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telephone</td>
<td></td>
<td>503-352-7317</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Study invitation, purpose, location, and dates
   You are invited to participate in a research study. In this study, we want to know whether a psychological test works the same for both males and females. This study has been approved by the Pacific University IRB and will be completed by August 2013. The study will take place at Coffee Creek Correctional Facility. The results will be used to inform mental health professionals about how this test works for men and women.

4. Participant characteristics and exclusionary criteria
You can participate in this study if you are 18 years or older and can read and speak English. You cannot participate if you are younger than 18 years old and cannot read or speak English. In addition, data is being used from two other research projects that studied similar topics. If you participated in one of these studies then you will not be able to participate in this study. The person who is conducting this study with you will go over questions with you to determine if you participated in one of the previous studies.

5. **Study materials and procedures**

You will be asked to complete two short surveys. One survey will ask you questions about yourself such as gender, age, race, marital status, and education level. After you complete the first survey, you will be given instructions on how to take the second survey. The second survey will be read aloud to you by the researcher, and consists of questions about different mental health symptoms.

About 200 other individuals will participate in the study. Participation will take about 15-20 minutes. It will not cost you anything to be a part of the study. If you do not wish to participate in the study, you will be free to return to your unit. A researcher will be present at all times to answer any questions you might have.

6. **Risks, risk reduction steps and clinical alternatives**

   a. **Unknown risks**

      It is possible that participation in this study may expose you to currently unforeseeable risks.

   b. **Anticipated risks and strategies to minimize/avoid**

      Some people may experience discomfort or slight anxiety by being asked to approach the test in a manner other then how they truly feel. If you begin to feel this way, you can talk to someone from Behavioral Health Services, or a staff member you trust.

   c. **Need for follow-up examination or care after the end of study participation**

      There is no anticipated follow up examination or care after participation has ended.

   d. **Advantageous clinical alternatives**

      This study does not involve experimental clinical trial(s).
7. Adverse event handling and reporting plan

The IRB office will be notified by the next normal business day if any adverse events occur. Should an adverse event occur, the investigator will locate an ODOC staff member to assist in contacting Behavioral Health Services. Only the information necessary to assist in reduction of the adverse event will be disseminated.

8. Direct benefits and/or payment to participants

It is important for you to understand that parole boards will not take into account your participation in this project in making decisions regarding your parole in any way.

a. Benefit(s)

There is no direct benefit to you as a study participant.

b. Payment(s) or reward(s)

You will not be paid for your participation.

9. Promise of privacy

The results of this study will be kept confidential. You will be assigned a random number that will be used instead of your name or State Identification Number (SID). This way no one can match your name to your responses except for the investigators. Your name and SID will be kept to monitor who has participated in the survey. That information will be kept on the principal investigator’s password protected computer. Your surveys will be kept in a locked case to be transported out of the facility. Once outside of the facility, your surveys will be kept in a locked filing cabinet in a locked room at Pacific University. After the data has been analyzed, all information with your name and SID will be destroyed. When we write or talk about what we learn from this study, we will leave things out so that no one will know we are talking about you.

While you are participating in this study, all rules and regulations of ODOC still count. If you tell the investigator of any danger to self or others, abuse of identifiable children, abuse of disabled or elderly persons, staff abuse of inmates, escape plans or attempts, or sexual assault, then ODOC staff will be notified. The Pacific University IRB will also be notified.
10. Medical care and compensation in the event of accidental injury

During your participation in this project it is important to understand that you are not a Pacific University clinic patient or client, nor will you be receiving complete mental health care as a result of your participation in this study. If you are injured during your participation in this study and it is not due to negligence by Pacific University, the researchers, or any organization associated with the research, you should not expect to receive compensation or medical care from Pacific University, the researchers, or any organization associated with the study.

11. Voluntary nature of the study

Your decision whether or not to participate will not affect your current or future relations with Pacific University or Oregon Department of Corrections. If you decide to participate, you are free to not answer any question or withdraw at any time without prejudice or negative consequences. You can choose to withdraw from the study up until you leave the interview room and return to your unit. If you choose to withdraw after beginning the study, your answers will not be used. We will keep what you have completed for 5 years in a locked filing cabinet.

12. Contacts and questions

The researcher(s) will be happy to answer any questions you may have at any time during the course of the study. If you are not satisfied with the answers you receive, please call Pacific University’s Institutional Review Board, at (503) 352-1478 to discuss your questions or concerns further. You will have to contact a staff member or your counselor in order to reach the Institutional Review Board. If you have questions about your rights as a research subject, or if you become injured in some way and feel it is related to your participation in this study, please contact the investigators and/or the IRB office. All concerns and questions will be kept in confidence.

13. Statement of consent

Yes    No
☐ ☐ I am 18 years of age or over.
☐ ☐ All my questions have been answered.
☐ ☐ I have read and understand the description of my participation duties.
☐ ☐ I have been offered a copy of this form to keep for my records.
☐ ☐ I agree to participate in this study and understand that I may withdraw at any time without consequence.

Participant’s signature ____________________________ Date ___________