ASSESSING DIFFERENTIAL ITEM FUNCTIONING ACROSS CLINICAL AND COMMUNITY SAMPLES IN THE MILLER FORENSIC ASSESSMENT OF SYMPTOMS TEST

A Dissertation

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

By

Mary Madison Eagle

May, 2015

ASSESSING DIFFERENTIAL ITEM FUNCTIONING ACROSS CLINICAL AND COMMUNITY SAMPLES IN THE MILLER FORENSIC ASSESSMENT OF SYMPTOMS TEST

Mary Madison Eagle, M.A.

APPROVED:

John P. Vincent, Ph.D.
Committee Chair

Gerald Harris, Ph.D.

Tonya Inman, Ph.D.

Ashley Stewart, Ph.D.
Department of Defense

Steven G. Craig, Ph.D.
Interim Dean, College of Liberal Arts and Social Sciences
Department of Economics

ASSESSING DIFFERENTIAL ITEM FUNCTIONING ACROSS CLINICAL AND COMMUNITY SAMPLES IN THE MILLER FORENSIC ASSESSMENT OF SYMPTOMS TEST

An Abstract of a Dissertation

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Doctor of Philosophy

By

Mary Madison Eagle

May, 2015

Abstract

Assessment of malingering is a central component of forensic evaluations in criminal as well as civil litigation contexts. As such, there is a need for empirically sound measures of malingering to accurately identify individuals who may be feigning symptoms for personal gain. The detection of differential item functioning (DIF) using item response theory methods provides a powerful method of evaluating whether items on a malingering assessment measure function differently in civil versus criminal litigation contexts. The aim of the current study was to evaluate DIF based on litigation type in the Miller Forensic Assessment of Symptoms Test (Miller, 2011). The M-FAST was administered to a sample of criminal defendants at Arkansas State Hospital as well as an analogue sample of civil litigants comprised of undergraduates at a large public Southwestern university. Results indicated DIF for nine M-FAST items, with five of the items more easily endorsed by criminal litigants and four of the items more easily endorsed by civil litigants. Implications of these results for research and clinical assessment are discussed.

Table of Contents

Background and Specific Aims

Assessment of malingering is a central component of forensic evaluations in criminal as well as civil litigation contexts. However, psychologists using clinical judgment alone to detect malingering perform only at chance levels and their findings often have little predictive value (Frederick & Crosby, 2000; Mossman, 2003). Moreover, criminal defendants have clear external incentives to alter their response styles during the course of a competency evaluation, such as the evasion of criminal prosecution (Heilbrun, 2001; American Psychiatric Association, 2000). Plaintiffs involved in civil litigation also often have incentives to appear impaired in these cases, such as the potential for financial gain. In light of this, forensic practitioners are advised to consider the possibility that clients undergoing evaluation will not present in a fully forthright manner.

The DSM –V (American Psychiatric Association, 2013) defines malingering as "the intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs." Annually, it is estimated 29% of personal injury cases, 30% of disability cases, 19% of criminal cases and 8% of medical cases involve probable malingering and symptom exaggeration (Mittenberg, Patton, Canyock, & Condit, 2002). There are significant costs associated with successful malingering. Indeed, false civil claims that are undetected have societal consequences such as increased insurance premiums and diversion of funds from the truly deserving to the undeserving (Bordini et al., 2002). Further, the administration of justice may be impeded where malingerers are not properly identified. Specifically, criminal charges may be dismissed and malingerers could achieve gains such as avoiding prison for treatment-based rehabilitation (Frederick, Crosby, &

Wynkoop, 2007). Consequently, there is a need for empirically sound measures of malingering to accurately identify individuals who may be feigning symptoms for personal gain (Frederick & Crosby, 2000).

To this end, the Miller Forensic Assessment of Symptoms Test (M-FAST; Miller, 2001) is often employed in forensic assessments, as it has been validated in studies with criminal defendants (Miller, 2001; Vitacco, Rogers, Gabel, &Munizza; 2007) as well as civil litigants (Guriel et al., 2004; Alwes et al., 2008) and can serve to measure response styles associated with symptom exaggeration or malingering. The M-FAST (Miller, 2001) is a 25-item structured interview designed to provide information regarding the probability that an individual is malingering. The M-FAST includes seven subscales, designed to measure differences in observed vs. reported symptoms; extreme symptomatology; rare combinations of symptoms; reports of unusual hallucinations; reports of unusual symptom course; reports of an unduly negative self-image; and suggestibility. The measure has a recommended cutoff score of 6 to differentiate between malingerers and honest responders (Miller, 2001).

However, the extent to which each individual M-FAST item measures an individual's malingering status and is able to discriminate between levels of malingering is unclear. An individual may obtain a score of 6 by responding affirmatively to numerous symptom patterns. As such, it is possible, even likely, that some M-FAST items provide more information regarding an individual's malingering status than others, and individuals responding affirmatively to these items may be more accurately identified as providing less than honest responses. Therefore, the total scores obtained by two individuals may reflect the same "level" of malingering, but may be comprised of items of varying degrees of discriminability. As the utility of individual M-FAST items for use within pre-trial forensic populations or simulated civil litigant populations has not yet been examined, an analysis of the specific functioning of individual M-FAST items may be of assistance in more accurately identifying individuals feigning psychopathology as well as in determining patterns and types of malingered responding.

Item Response Theory

Existing data on malingering measures, including the M-FAST (Miller, 2001), has traditionally been derived through the traditional classical test theory (CTT) framework (Lord &Novick, 1968). However, when evaluating measure utility and item functioning, Item Response Theory (IRT) offers advantages over CTT (Embretson & Reise, 2000). IRT, broadly, is a technique used to establish the psychometric properties of items and scales. It improves upon CTT in three primary ways. First, as opposed to CTT which uses the sum of various items to represent a trait, IRT obtains a trait score for each item. Second, IRT provides the reliability of each item at different levels of the underlying construct, whereas CTT offers only reliability for the measure as a whole. Third, CTT psychometric properties are sample dependent, and thus, vary across samples, but IRT psychometric properties are assumed to be sample independent.

IRT refers to a test conceptualization approach that examines the relationship between an individual's item response, and an underlying latent trait, commonly referred to as theta (Thomas, 2011; Gray-Little, Williams, & Hancock, 1997). Theta is generally presented on a z-score scale, thus making interpretations intuitively appealing. There are two assumptions that must be met within IRT. First, the test must be unidimensional, meaning that only a single latent trait is being measured. Second, the test must display local independence, such that the latent trait is accounting for all covariation between the items (Thomas, 2011; Gray-Little, Williams, & Hancock, 1997).

IRT models generally include two parameters, the threshold parameter (*b*), and the discrimination parameter (*a*). The threshold parameter, also referred to as the difficulty parameter, serves to identify the point along the latent trait continuum at which respondents have a 50 percent chance of endorsing an item. The discrimination parameter, signifies the relationship between the item and theta, and provides information regarding the extent to which the item discriminates among individuals with different levels of the underlying trait (Thomas, 2011; Fraley, Waller, & Brennan, 2000; Gray-Little, Williams, & Hancock, 1997).

The relationship between the item threshold parameter and the discrimination parameter is displayed by Item Characteristic Curves (ICC) or trace lines. The graphs of trace lines include the underlying trait of interest (in this case, malingering) on the x-axis, and the probability of endorsement on the y-axis, and the trace lines depict the probability of a given response at different levels of theta. The slope of the trace line is the discrimination parameter, and the larger this value is, the better the item is able to discriminate among individuals at different levels of theta (Thomas, 2011; Gray-Little, Williams, & Hancock, 1997). Item Information Curves (IIC) can also be derived to display the amount of information provided by the item at a given trait level. Items are most informative at the difficulty parameter, and items with large discrimination parameters provide more information than those with smaller discrimination parameters (Fraley, Waller, & Brennan, 2000). Importantly, the information from these item-level parameters and curves can be combined to provide test-level data.

The Test Information Curve (TIC) is the sum of all Item Information Curves, and represents the relative precision of the scale across different levels of theta. The Test Information Curve is the inverse of the standard error of measurement for the test at different levels of theta.

Thus, in IRT there is a standard error for each level of theta measured, and not a single standard error for the entire test, as is the case in CTT. The Test Characteristic Curve (TCC) is the expected summed raw score for each value of theta, and depicts the nonlinear relationship between the raw scores on the scale and theta (Fraley, Waller, & Brennan, 2000; Gray-Little, Williams, & Hancock, 1997).

Functions of IRT

Perhaps the most well-recognized application of IRT has been the utilization of this framework within standardized testing, resulting in what is now known as Computer Adapted Testing. This approach allows tests to be shorter in length, while maintaining or improving upon reliability and standardization. IRT has also been applied to numerous psychological tests of personality and psychopathology.

A second key application of IRT is that it provides an attractive framework for investigating the degree to which items differ or are invariant across samples (Embretson & Reise, 2000). By modeling the trait-item relationship, IRT can characterize differences in item functioning in a way that is not impacted by differences in trait distributions across the two populations being compared (Embretson & Reise, 2000). An item is said to exhibit differential item functioning (DIF) when the trait-item relationship is found to be different across populations.

Differential Item Functioning

Differential Item Functioning (DIF), also known as measurement bias, occurs when different groups with the same latent trait (e.g., false responding) have a different probability of giving a certain response on an assessment measure. The main goal for DIF analyses is to identify items on an assessment measure that may assess traits differently in different samples. Measurement bias can be assessed for diverse participants in the same sample (e.g. do items function differently for Caucasians and African Americans on a measure of personality pathology) or for participants in two different samples (e.g., do items function differently for civil and criminal litigants on a measure of malingering) as is the case in the current study.

If a measure consists of dichotomously scored items, such as those items on the M-FAST, then uniform or non-uniform DIF may exist. Uniform DIF occurs when the magnitude of conditional dependency is relatively invariant across the latent trait continuum (Walker, 2011). In other words, when uniform DIF is present, an item of interest is consistently more likely to be endorsed (or not endorsed) for one group over another. This is contrasted with non-uniform DIF, which occurs when a shift in the likelihood of item endorsement is not consistent across the latent trait continuum.

DIF analysis is a micro-statistical procedure. That is, DIF analyses attempt to identify performance differences on individual items as opposed to performance differences on an overall measure (Walker, 2011). DIF is found by examining differences in item characteristic curves across the groups of interest. DIF analyses are important in the test validation process to ensure that scores obtained from psychological measures are unbiased and reflect the same construct for all respondents (Walker, 2011). As the M-FAST is considered to be a brief, valid, and reliable screening measure of malingering, it is frequently used to detect false-responding in both criminal and civil litigant populations. Because of the high base rates of malingering in both civil and criminal populations in addition to the extensive real-world implications of failing to accurately classify false-responders, examining the M-FAST for equivalent functioning across samples of both criminal and civil litigants is of importance to clinicians and researchers alike.

Current Study

The current study had two specific aims. First, we sought to apply IRT principles to analyze the psychometric properties of the M-FAST. Specifically, the study attempted to examine the properties of the M-FAST from the IRT framework in two populations. By examining this measure from an IRT perspective, it is possible to evaluate the extent to which items are related to other items and to the underlying trait of malingering, the extent to which the items provide information about malingering within pretrial criminal defendants as well as civil litigants, and the extent to which these items are able to discriminate between individuals at various levels of the latent trait of malingering. Second, the presence of differential item functioning on item level data of the M-FAST in criminal and (simulated) civil litigant populations was tested. By investigating the cumulative effects of DIF across civil and criminal litigant populations, researchers can ascertain whether the same measure (M-FAST) as applied to adults in civil litigation and criminal litigation contexts represents different latent levels of malingering in these two populations, implying scalar equivalence or inequivalence.

Methods

Participants

Criminal Litigant Sample. This sample (N= 72) was obtained utilizing a retrospective review of the records of forensic patients at Arkansas State Hospital (ASH) who underwent court-ordered evaluations of their competency to proceed to trial and/or an evaluation of their mental state at the time of their offense. These patients were housed either within ASH or were evaluated at ASH on an outpatient basis. Participants included those patients who were aged 18 to 65 at the time of their evaluation, who presented to ASH for a forensic evaluation, and who were administered the M-FAST during the time period between January 2002 and July 2013.

Civil Litigant Sample. This analogue sample is part of a larger research initiative in which 523 undergraduates enrolled in psychology courses at a large public Southwestern University were recruited to participate. Participants earned extra credit for their participation. Inclusion criteria required that the participant be older than the age of 18 and hold a valid driver's license so that all participants could presumably experience a situation similar to the one depicted. The experimental manipulation asked participants to complete the M-FAST as if they had been involved in a motor vehicle accident under different instructional conditions that varied possible incentives to over-report emotional symptoms.

Measure

Miller Forensic Assessment of Symptoms Test (Miller, 2001). The M-FAST is a 25-item structured interview designed to provide information regarding the probability that an individual is malingering. The M-FAST includes seven subscales, designed to measure differences in observed vs. reported symptoms; extreme symptomatology; rare combinations of symptoms; reports of unusual hallucinations; reports of unusual symptom course; reports of an unduly negative self-image; and suggestibility. The measure has a recommended cutoff score of 6 to differentiate between malingerers and honest responders (Miller, 2001). The M-FAST has been standardized and validated on both clinical and non-clinical samples, with total score reliability estimates of .93 and .92, respectively. Average correlation of individual items with the total score ranged from .35-.85, and subscale alphas ranged from .61 to .81 (Miller, 2001). The interrater reliability of the M-FAST has been shown to be better than 99 percent. Using both simulation and known-group designs, Miller demonstrated the criterion, convergent, and discriminant validity of the M-FAST over 4 studies by comparing the M-FAST to other assessments such as the SIRS (Rogers, 1997), the MMPI-2 validity scales, and the M-Test (Miller, 2001). The validity of the M-FAST generalized across race, gender, age, and setting (Miller, 2001). The measure has shown sensitivity of .93 and specificity of .83 at this cutoff score in clinical samples.

Procedure

Criminal Litigant Sample.

Data were identified and collected from individual forensic evaluators and medical records by reviewing the patient charts for appropriate inclusion criteria. All study data was obtained from the forensic medical files of patients who were evaluated at ASH. Individuals who underwent a forensic evaluation and were administered the M-FAST were identified by individual forensic evaluators, and relevant data was obtained from patient files.

Civil Litigant Sample.

After participants gave informed consent to participate in the study, they were asked to read one of four instructional scenarios (no-litigation, post-litigation, active-litigation with no suggestion to malinger, or active-litigation including a suggestion that the more severe symptoms they reported the greater the potential monetary gain) and were administered the M-FAST interview by a trained research assistant. Demographic characteristics assessed included age, ethnicity, gender, education, work status, occupation, and history of mental health services. The

Trauma Symptom Inventory (TSI; Briere, 1995), the Traumatic Experiences Checklist (TEC; Nijenhuis, Van der Hart, & Kruger, 2002), and the Life Experiences Survey (LES; Sarason, Johnson, & Siegel, 1978) were also administered in this sample, but not included in the current analyses.

Instructional Conditions. The instructional conditions under which participants responded asked them to imagine that they had been in a motor vehicle accident with truck owned by a major wholesaler in which they had sustained no serious physical injury, but were still experiencing some emotional difficulties relating to the accident including jumpiness/nervousness while driving, avoidance of the location of the accident, avoiding conversations about the accident, having bad dreams about the accident, and experiencing an exaggerated startle response. Participants were then informed that they either 1) were content with their settlement from their insurance company, but had been asked by their physician to see a psychologist who administered psychological measures to determine the level of their impairment (no litigation), 2) were content with the outcome of their lawsuit, but had been asked by their physician to see a psychologist who administered psychological measures to determine the level of their impairment (post-litigation), 3) were currently in-litigation and were told by their legal counsel that the wholesaler's lawyers requested that they see a psychologist who administered psychological measures to determine the level of their impairment (active litigation with no suggestion to malinger), or 4) were currently in-litigation and were told by their legal counsel that the wholesaler's lawyers requested that they see a psychologist who administered psychological measures to determine the level of their impairment, and had been informed prior to completing the measures that the more impaired they appeared, the more monetary damages they likely would be awarded (active litigation including a suggestion to malinger). The data for this sample were collected in a prior study that examined the impact of the instructional conditions on M-FAST scores and reported trauma symptoms (Christiansen & Vincent, 2012).

Data Analytic Plan

For the purposes of this study, data from the simulated civil litigation sample were only utilized from participants in the two “active litigation” conditions (i.e. conditions numbered (3) and (4) from the scenario described above) (N = 254). Analyzing the data in this way preserved internal validity, as data from the criminal litigant sample was gathered only from participants involved in active litigation. Moreover, as noted by Gutheil (2003), it is often the case that attorneys will offer their clients a suggestion to exaggerate their symptoms in order to maximize their chances to obtain an external incentive or mitigate their punishment in some way, in the case of criminal litigation. Therefore, it also made logical sense for the purposes of increasing external validity to examine differential item functioning of the M-FAST for participants involved in active litigation in both criminal and analogue samples.

The IRT model fitting and the computation of test statistics were performed using IRTPRO (Cai, du Toit, and Thissen, 2011). Goodness of fit of the models was evaluated using the M_2 statistics and its associated RMSEA value (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005, 2006; Thissen, 2009), as well as the standardized local dependence (LD) chi-square indices (based on the LD index proposed by Chen & Thissen, 1997). The M_2 statistic represents a suitable alternative for G^2 when the table of item response patterns becomes too sparse to compute the likelihood goodness-of-fit chi-square statistic, as is the case in our current study. Local dependence indicates that the observed covariation among responses to the items in an item-pair exceeds that predicted by the model. The LD indices are standardized chi-square values; values 10 or greater are considered noteworthy (Thissen, 2009) and thus challenge the assumption of unidimensionality.

The 2PL model was fitted to the 25 dichotomously scored M-FAST items. For the purposes of the present study, the 2PL model represents the probability of endorsing an M-FAST item as a function of the underlying construct of malingering. For each item, two types of parameters are estimated – *discrimination* (or slope) and *threshold* (Embretson & Reise, 2000). The discrimination parameter represents the degree of association between the item response and the underlying construct. The threshold parameter provides information regarding the extent to which the item discriminates among individuals with different levels of the underlying trait.

The presence of DIF was investigated using the approach advanced by Thissen, Steinberg, and Wainer (1993). Differences in parameter estimates between groups are evaluated using model comparison tests. To implement this approach, a subtest of items (“anchor items”) is identified as a means to “link” the groups (allowing for an estimated population mean group difference in the underlying construct). Edelen et al., 2006 recommend identifying anchor items by using an exploratory, iterative process whereby each item is initially tested for DIF by using all other items as the anchor set. Items not showing DIF at this step are regarded as anchor items; the remaining items, referred to as the ‘studied’ or ‘candidate’ items, are then evaluated for DIF. Wald tests based on the procedure proposed by Lord (1977), providing separate chi-square statistics for the discrimination and threshold parameters for each studied item are used to evaluate for the presence of DIF. When DIF is detected, effect size for the threshold and/or slope parameters will aid the description and interpretation of the group differences (Steinberg & Thissen, 2006).

Results

Item Response Theory Analyses

Unidimensionality. In an analysis of the item response theory data for civil litigants, the 2PL unidimensional IRT model showed satisfactory fit: $M_2(275) = 399.08$, $p = .0001$; *RMSEA* = .04. The significant M_2 statistic indicates some model error; however, the *RMSEA* indicates acceptable fit of the model. Unfortunately, due to insufficient sample size, the 2PL unimdimensional IRT model fit could not be assessed for the criminal litigant population. Analyses for criminal litigant group were carried out under the assumption that the unidimensional model adequately fit the data. None of the standardized LD statistics approached the value of 10.0 for either the civil or criminal litigant populations. For civil litigants, the largest LD value was observed between item 2 (I feel depressed most of the time) and item 24 (On many days I feel so bad that I can't even remember my full name). LD $\chi^2 = 8.3$. For criminal litigants, the largest LD value was observed between item 11(Whenever I am sitting in a chair, I have to breathe deep breaths in order not to get sick) and item 16 (Sometimes I am convinced that I have more than one personality). LD $\chi^2 = 3.8$. These findings with respect to unidimensionality and local dependence offered justification for proceeding with unidimensional IRT analyses. As stated previously, the results of all analyses conducted within the criminal litigant population should be interpreted with caution, given the small sample size.

Detection of DIF. The first step in conducting the DIF analyses was to identify a set of anchor items for linking the civil and criminal litigant groups. To do so, each item was initially tested for DIF using all the other items as a tentative anchor. 16 items emerged as not exhibiting DIF, as evidenced by non significant Wald (χ^2) statistics ($p > .05$). Items with significant Wald

(χ^2) statistics are as follows: Item 2 (I feel depressed most of the time; p = .005); Item 4 (Do voices tell you to do things and, if yes, do you obey them?; p = .004); Item 6 (I experience hallucinations that last continually for days; p = .009); Item 10 (Most times when people are talking to me, I see the words they speak spelled out; p = .033); Item 11 (Whenever I am sitting in a chair, I have to breath deep breaths in order not to get sick; p > .001); Item 12 (Some nights I have nightmares so bad it scares me and, if yes, does this only happen when you have lost a lot of weight?; p = .021); Item 13 (Lately my eyesight is so good that I think I have a special power; p = .031); Item 21 (Sometimes I hear music coming from nowhere; p = .029); and Item 23 (Most of the time I feel that I don't really matter; p > .001).

In a separate analysis, the remaining 16 items were evaluated for DIF to confirm their appropriateness in serving as anchor items. None of the Wald statistics approached significance, indicating a suitable anchor set. The remaining nine items constituted the study items and were evaluated for DIF using the 16-item anchor.

For evaluating the Wald tests for the nine studied items, Type I error rate was controlled using the Benjamin-Hochberg (B-H, 1995) multiple comparisons procedure. All nine of the studied items exhibited DIF. In each case (with the exception of item 13, χ^2 (1) = 2.7, p = .104), the DIF was concentrated in the threshold parameter, as evidenced by a significant Wald test statistic: for item 2, χ^2 (1) = 5.0, p = .025; for item 4, χ^2 (1) = 9.0, p = .002; for item 6, χ^2 (1) = 6.3, p = .012; for item 10, χ^2 (1) = 5.6, p = .018; for item 11, χ^2 (1) = 25.8, p < .001; for item 12, χ^2 (1) = 7.2, p = .007; for item 21, χ^2 (1) = 5.7, p = .017; for item 23, χ^2 (1) = 32.6, p <.001). Wald test statistics for slope parameters for seven of the nine items were nonsignificant: for item 4, χ^2 (1) = .60, p = .446; for item 6, χ^2 (1) = 1.0, p = .306; for item 10, χ^2 (1) = .2, p = .625; for item 11, χ^2 (1) = 3.4, p = .065; for item 12, χ^2 (1) = .10, p = .726; for item 21, χ^2 (1) = 1.1, p = .284; and for item 23, χ^2 (1) = .10, p = .733. Item 2, χ^2 (1) = 5.4, p = .020; and item 13, χ^2 (1) = 5.2, p = .022 displayed significant Wald test statistics for slope parameters.

A final calibration of item parameters was performed by fitting a model in which the slope and threshold parameters for the anchor items were constrained to be equal across the civil and criminal litigant groups and the slope parameters for the nine study items were constrained to be equal across civil and criminal litigant groups, with the threshold parameters freely estimated (see Table 1). Goodness of fit for this model was acceptable: $M_2(264) = 399.41$, $p < .001$; *RMSEA* = .04. The significant M_2 statistic indicates some model error; however, the *RMSEA* indicates acceptable fit of the model. Of note, these goodness of fit data should be interpreted with caution, as the sample size in the criminal litigant population is not sufficient to calculate a reliable value of the M_2 statistic.

DIF Items. As mentioned, eight of the nine study items showed significant DIF in the threshold parameters across the civil and criminal litigant groups. These item parameters are presented in Table 1. In the case of items 2 (I feel depressed most of the time), 4 (Do voices tell you to do things and, if yes, do you obey them?), 6 (I experience hallucinations that last continually for days), 13 (Lately my eyesight is so good that I think I have a special power), and 23 (Most of the time I feel that I don’t really matter), the direction of DIF was such that it was “easier” for criminal litigants to endorse the M-FAST item. In the case of items 10 (Most times when people are talking to me, I see the words they speak spelled out), 11 (Whenever I am sitting in a chair, I need to breathe deep breaths in order not to get sick), 12 (Some nights I have nightmares so bad it scares me and, if yes, does this only happen when you have lost a lot of weight?), and 21 (Sometimes I hear music coming from nowhere), the direction of the DIF was such that it was “easier” for civil litigants to endorse the M-FAST item. Figure 1 shows the trace lines for all nine items. As depicted, the lines differ in their left-to-right locations (i.e., threshold parameters) for civil and criminal litigants. Figure 2 shows the test characteristic and test information curves by litigation type.

Anchor Items. The 2PL results showed that all M-FAST items were found to be adequately discriminating, with the exception of item 5 (*a* = .46). Discrimination (slope) parameters are analogous to factor loadings in traditional or confirmatory factor analysis. Discrimination parameters can in fact be translated to factor loadings (McLeod, Swygert, & Thissen, 2001, p. 199). Values that are 1.0 (corresponding to a factor loading of .50) or greater are considered substantial. The discrimination parameters (with the exception of item 5) ranged from 1.00 (Item 1: I often find myself not being able to sit in a chair) to 4.18 (Item 19: I often get the strange feeling that I am from another planet). Threshold parameters were generally located above the mean (with the exception of items 1, 3, 14, 20, and 22) ranging from .55 (Item 17: The times when you can’t go to sleep, do you often smell strange odors that are not really there?) to 2.41(Item 5: I feel unusually happy most of the time). Threshold parameters that fell below the mean ranged from -.25 (Item 1: I often find myself not being able to sit in a chair) to .39 (Item 22: When I hear voices, I often develop fears of leaving my house or room).

Discussion

The main aim of the current study was to examine the measurement equivalence (or presence of DIF) of each of the M-FAST items across civil and criminal litigant populations using IRT. The rationale for this study lies in the dearth of studies examining M-FAST item functioning, despite the measure being frequently used for detection of malingered response styles in both civil and criminal litigant populations. To this writer's knowledge, this is the first study to use IRT to evaluate DIF across criminal and civil litigant populations. Widiger and Spitzer (2001) suggested that bias can function at two levels: assessment bias (e.g., a biased application of the diagnostic criteria for malingering) and criterion bias (e.g., bias within the defining criteria for malingering).

Studies of malingering presence in litigation cite differing prevalence rates across litigation type: 29% of personal injury cases, 30% of disability cases, 19% of criminal cases and 8% of medical cases (Mittenberg, Patton, Canyock, & Condit, 2002). The results of DIF analyses in the current study suggested invariance across litigation type for 16 of the 25 M-FAST items. Nine of the items, however, functioned differently across litigation types, suggesting that it is "easier" for criminal litigants to endorse the following items: "I feel depressed most of the time," "Do voices tell you to do things and, if yes, do you obey them?," "I experience hallucinations that last continually for days," "Lately my eyesight is so good that I think I have a special power," and "Most of the time I feel that I don't really matter," whereas civil litigants are more "easily" endorse the following: "Most times when people are talking to me, I see the words they speak spelled out," "Whenever I am sitting in a chair, I need to breathe deep breaths in order not to get sick," "Some nights I have nightmares so bad it scares me and, if yes, does this only happen when you have lost a lot of weight?," and "Sometimes I hear music coming from nowhere."

It is unclear whether the aforementioned varying prevalence rates of malingering in different types of litigation reflect bias in the measure used to assess malingering or criterion bias. In the present study, 36% of the items on the M-FAST exhibited DIF and those items exhibiting DIF were essentially equally distributed in terms of the number of items more easily endorsed by criminal litigants (5 items) and the number of items more easily endorsed by civil litigants (4 items). These results suggest that while some items on the measure may be biased, the measure does not appear to universally "favor" one litigant group over another in terms of ease of item endorsement.

When considering possible explanations for DIF, it is important to bear in mind that DIF may result from "bias" in the traditional sense, in which there is a problem in the wording of a given test item such that it favors members of a particular group (Sharp et al., 2014). This is known as a measurement artifact (Michonski, 2011; Wicherts & Dolan, 2010). On face value, the M-FAST items that showed DIF in the current study do not appear thematically to display bias toward one type of litigants over another. That is, the items do not seem to contain content that would make one litigant group more likely to endorse the item than another litigant group. Further, given that the M-FAST response options are True/False, scoring of the items is not susceptible to variations based on the clinical judgment of the particular examiner. Therefore, it appears that, in the current study, the biases do not seem to lie at the level of the assessment instrument.

Additionally, measurements artifacts may occur for reasons of social desirability. An example of this would be if civil litigants were less likely to endorse the item "I experience hallucinations that last continually for days" because they believe it makes them appear "crazy." In this vein, civil litigants may be willing to endorse less stigmatized mental health symptoms such as depression, but unwilling to endorse mental health symptoms (e.g., psychosis) that may carry more of a negative stigma. Alternatively, DIF may simply occur because, in addition to the common factor that is being measured, a given item taps into a specific factor that truly does differ among subgroups (Michonski, 2011; Wicherts & Dolan, 2010).

Several limitations in the current study should be considered. The most important limitation to note is the small sample size in the criminal litigant population. Given that a sample size of 72 falls well below the recommended sample size of 300 for conducting IRT analyses with sufficient power to assess item performance, all results put forth in the current study should be interpreted with caution. Second, demographic data was unavailable for the criminal litigant population due to feasibility issues. Because the current standard is for a “panel of experts” to convene to decide on the sources of DIF within a measure (i.e., to determine whether DIF is a sign of bias or not), varying explanations can exist as to the source of DIF. As the process is subjective, there is often no agreement about the true source of the DIF. Therefore, additional demographic data may have been helpful in providing additional possible explanations for the presence of DIF on the M-FAST items. Finally, it should be noted that we are comparing a simulated civil litigant population with a population of individuals who are truly involved in criminal litigation. This may create issues for external validity, as incentive to alter response styles is certainly stronger if the anticipated gains are tangible rather than hypothetical.

Taken together, the findings of the current study provide evidence for the notion that DIF exists within the M-FAST. However, because the nature of these findings is tenuous due to sample size limitations, any further interpretation as to the source of the DIF is cautioned until these results are replicated (or refuted) with a larger sample of criminal litigants.

Future Directions. There are many different ways to approach DIF analyses, which can range from traditional item-difficulty based approaches to Item Response Theory to Structural Equation modeling. The issue is that these can oftentimes produce different results (Bachman, 2004; McNamara & Roever, 2006). Further, variations of the same analytical approach can lead to different results based on group characteristics (e.g., differing sample sizes, focal group ability distribution) or different approaches to interpreting results (e.g., varying ideas on what p-value determines statistical significance in DIF analyses) (Karami & Salmani, 2011). Therefore, future research should focus on testing for the presence of DIF on the M-FAST in a larger sample of civil and criminal litigant populations using a variety of analytic methodologies.

Table 1

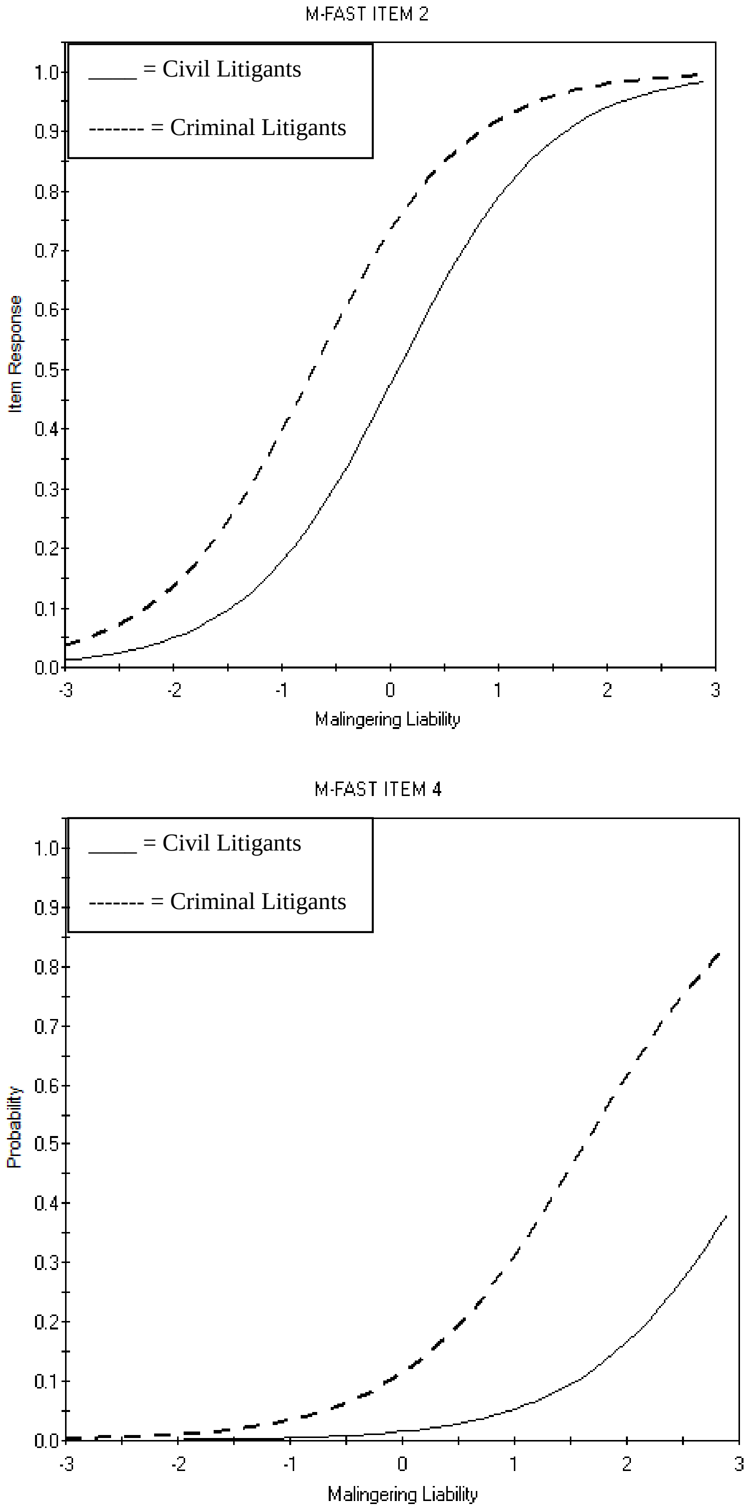
IRT Item Parameter Estimates

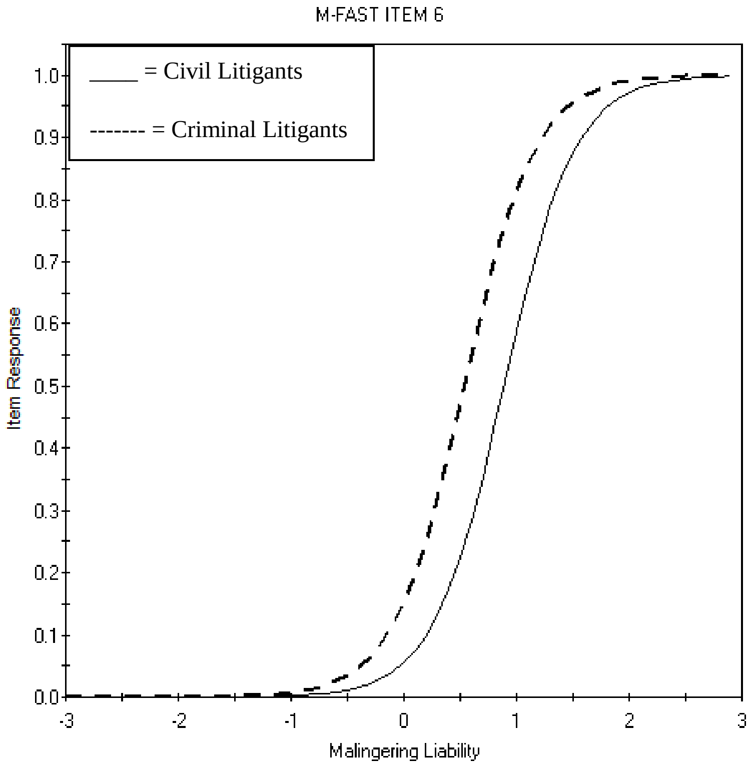
Item	Litigation Type	*a*	*b*
Anchor Items			
1.	Both	1.00 (.23)	-.25 (.13)
3.	Both	1.38 (.26)	.33 (.12)
5.	Both	.46 (.22)	2.41 (1.17)
7.	Both	1.78 (.35)	1.07 (.19)
8.	Both	3.06 (.52)	.73(.11)
9.	Both	4.09 (.76)	.82 (.10)
14.	Both	2.71 (.42)	.03 (.07)
15.	Both	2.53 (.42)	.64 (.11)
16.	Both	2.52 (.41)	.58 (.10)
17.	Both	2.66 (.43)	.55 (.10)
18.	Both	2.71 (.44)	.59 (.10)
19.	Both	4.18 (.90)	1.16 (.14)
20.	Both	2.83 (.44)	.27 (.08)
22.	Both	3.80 (.63)	.39 (.08)
24.	Both	2.67 (.45)	.79 (.12)
25.	Both	2.06 (.39)	1.08 (.18)
Threshold DIF			
2.	Civil Litigants	1.43 (.28)	.07 (.12)
	Criminal Litigants	1.43 (.28)	-.71 (.24)
4.	Civil Litigants	1.26 (.54)	3.28 (1.39)
	Criminal Litigants	1.26 (.54)	1.62 (.47)
6.	Civil Litigants	3.20 (.73)	.89 (.16)
	Criminal Litigants	3.20 (.73)	.54 (.14)
10.	Civil Litigants	1.88 (.56)	1.01 (.28)
	Criminal Litigants	1.88 (.56)	1.95 (.38)
11.	Civil Litigants	2.12 (.41)	.20 (.10)
	Criminal Litigants	2.12 (.41)	1.54 (.27)

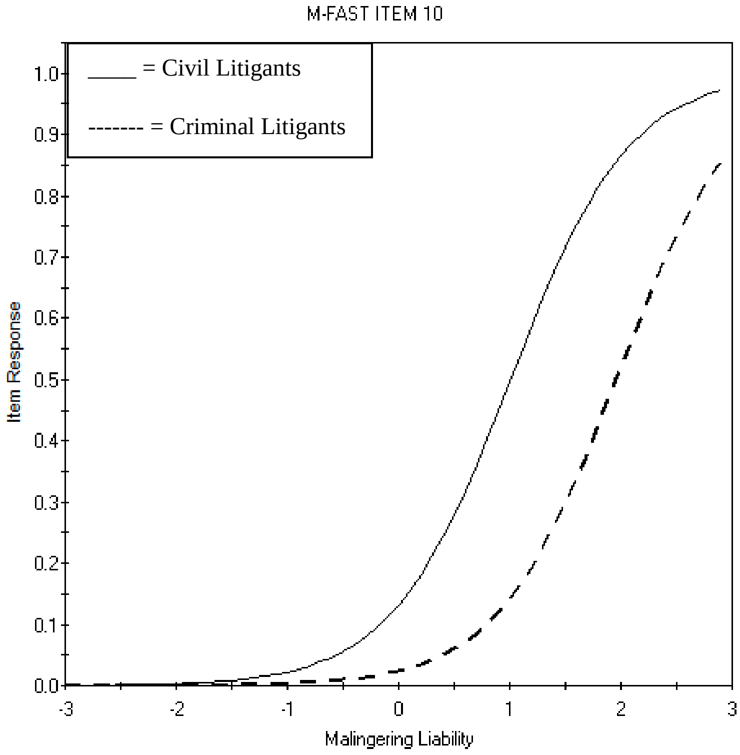
12.	Civil Litigants	1.79 (.37)	.63 (.16)
	Criminal Litigants	1.79 (.37)	1.40 (.28)
13.	Civil Litigants	1.27 (.57)	2.80 (1.22)
	Criminal Litigants	1.27 (.57)	.03 (.07)
21.	Civil Litigants	3.40 (.54)	.23 (.08)
	Criminal Litigants	3.40 (.52)	.62 (.14)
23.	Civil Litigants	2.04 (.38)	.62 (.14)
	Criminal Litigants	2.04 (.38)	-.64 (.19)

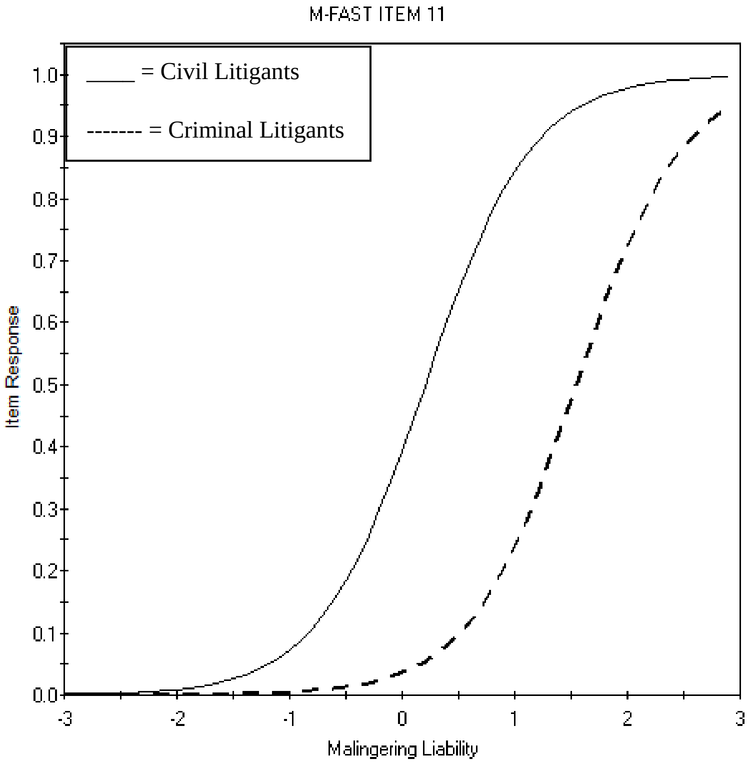
Note. Values enclosed in parentheses represent standard errors of the parameter estimates, *a* represents the discrimination parameter, *b* represents the threshold parameter. For items 2, 4, 6, 10, 11, 12, 13, 21, and 23, the slope parameter estimates have been constrained equal for Civil Litigants and Criminal Litigants and the threshold parameter estimates are estimated separately for Civil Litigants and Criminal Litigants.

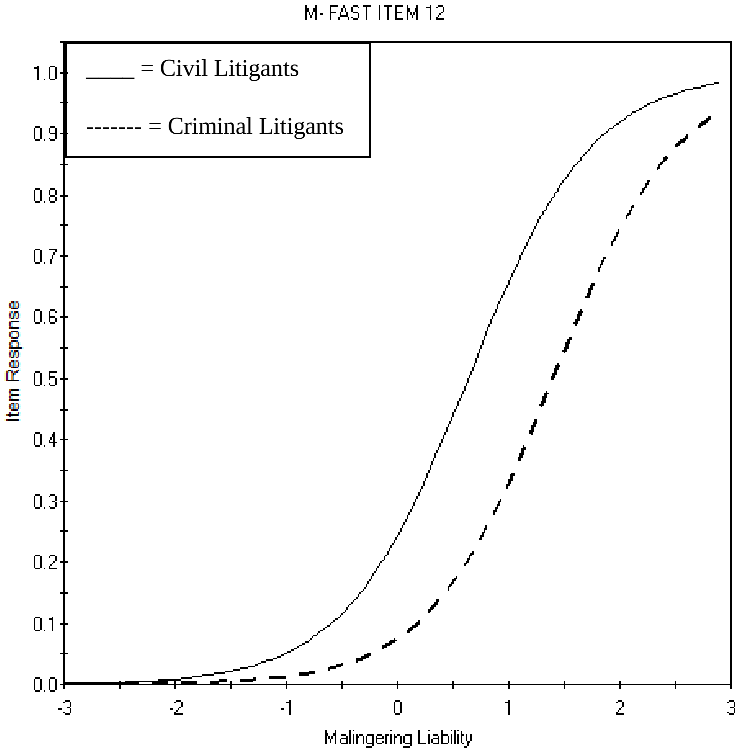
Figure 1. Item Characteristic Curves for 9 M-FAST items showing DIF. These curves depict differential item functioning with respect to thresholds across litigation type.

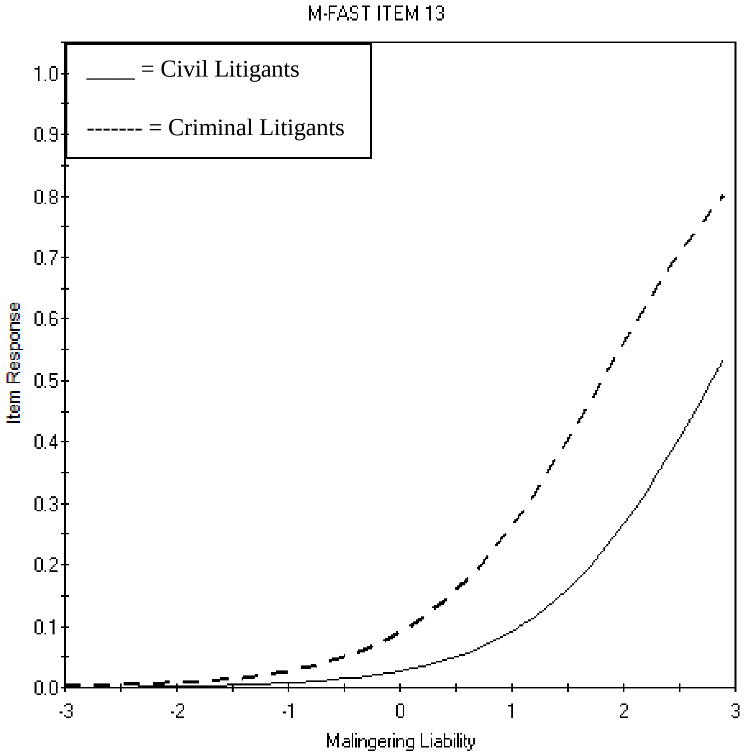


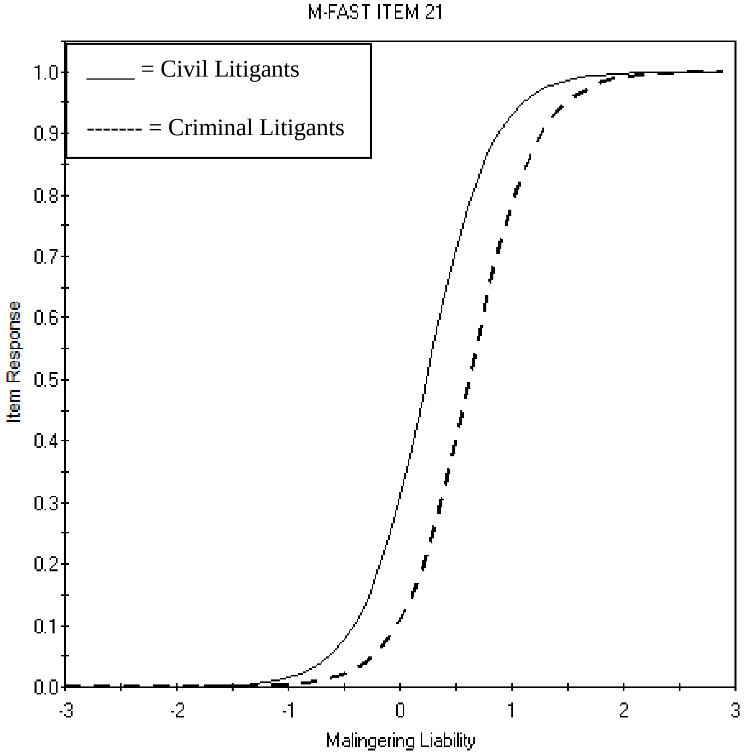

M-FAST ITEM 6
_____ = Civil Litigants
------- = Criminal Litigants
Item Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Malingering Liability


M-FAST ITEM 10
_____ = Civil Litigants
------- = Criminal Litigants
Item Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Malingering Liability


M-FAST ITEM 11
_____ = Civil Litigants
------- = Criminal Litigants
Item Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Malingering Liability


M- FAST ITEM 12
_____ = Civil Litigants
------- = Criminal Litigants
Item Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Malingering Liability


M-FAST ITEM 13
_____ = Civil Litigants
------- = Criminal Litigants
Item Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Malingering Liability


M-FAST ITEM 21
_____ = Civil Litigants
------- = Criminal Litigants
Item Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Malingering Liability

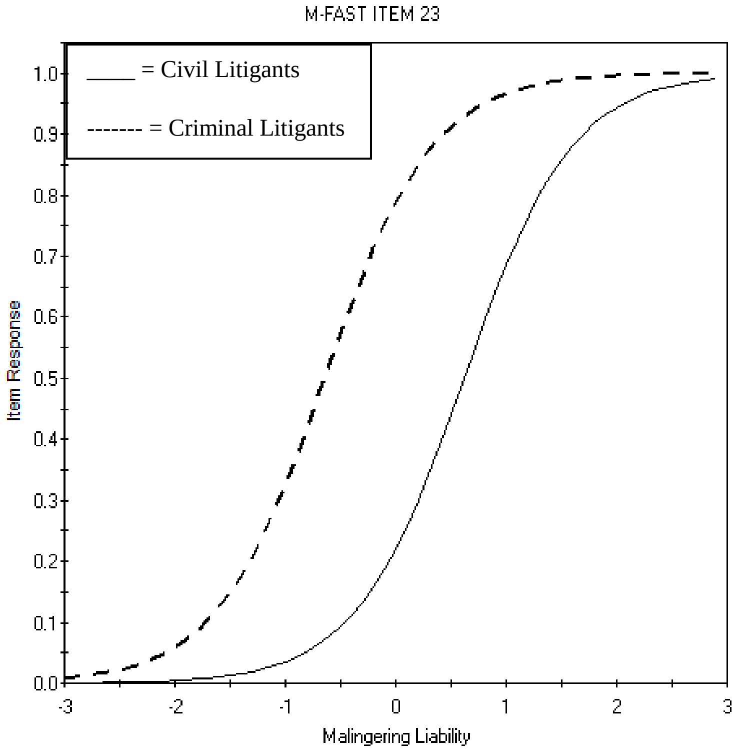
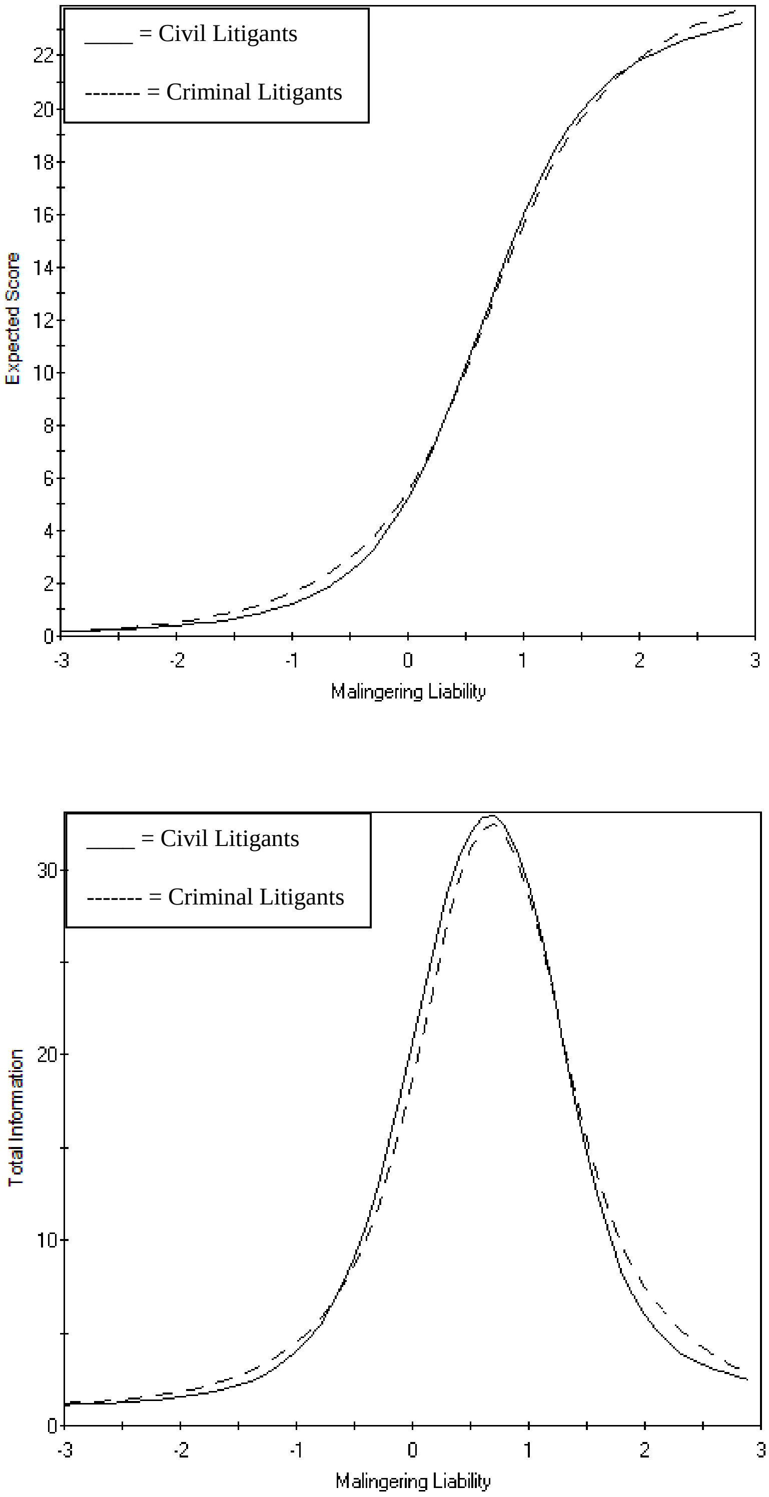

M-FAST ITEM 23
_____ = Civil Litigants
------- = Criminal Litigants
Item Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Malingering Liability

Figure 2. Test Characteristic and Test Information Curves (from top to bottom) by litigant type



References

Alwes, Y. R., Clark, J. A., Berry, D. R., & Granacher, R. P. (2008). Screening for feigning in a civil forensic setting. *Journal Of Clinical And Experimental Neuropsychology*, *30*(2), 1-8.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text Revision). Washington, DC.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bordini, E. J., Chaknis, M. M., Ekman-Turner, R. M., & Perna, R. B. (2002). Advances and issues in the diagnostic differential of malingering versus brain injury. *Neurorehabilitation, 17*(2), 93-104.

Briere, J. (1995). *The Trauma Symptom Inventory professional manual*. Odessa, FL : Psychological Assessment Resources, Inc.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible professional item response theory modeling for patient reported outcomes [Computer software]. Chicago, IL: SSI International.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2 tables. The British Journal Of Mathematical And Statistical Psychology, 59(Pt 1), 173-194.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22, 265–289.

Christiansen, A. K. & Vincent, J. P. (2012). Assessment of litigation context, suggestion, and malingering measures among simulated personal injury litigants. *Journal of Forensic Psychology Practice*, *12*(3), 238-258.

Denney, R. (2007). Assessment of malingering in criminal forensic neuropsychological settings. In K. Boone (Ed.), *Assessment of feigned cognitive impairment: A neuropsychological perspective* (pp. 428-452).

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach - Application to the Mini-Mental State Examination. Medical Care, 44, S134–S142. doi:10.1097/01.mlr.0000245251.83359.8c

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Fraley, R., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality And Social Psychology*, *78*(2), 350-365.

Frederick, R. I., & Crosby, R. D. (2000). Development and validation of the Validity Indicator Profile. *Law and Human Behavior, 24*(1), 59-82.

Gray-Little, B., Williams, V. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *23*(5), 443-451.

Guriel, J., Yañez, T., Fremouw, W., Shreve-Neiger, A., Ware, L., Filcheck, H., & Farr, C. (2004). Impact of Coaching on Malingered Posttraumatic Stress Symptoms on the M-FAST and the TSI. *Journal of Forensic Psychology Practice, 4*(2), 37-56.

Gutheil, T. G. (2003). Reflections on Coaching by Attorneys. *Journal of The American Academy Of Psychiatry And The Law, 31*(1), 6-9.

Heilbrun, K. & Kramer, G. (2001). Update on risk assessment in mentally disordered populations. *Journal of Forensic Psychology Practice, 1*(2), 55-63.

Karami, H., & Salmani Nodoushan, M. A. (2011). Differential Item Functioning: Current Problems and Future Directions. International Journal Of Language Studies, 5(3), 133-142.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford England: Addison-Wesley.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Portinga (Ed.), Basic problems in cross-cultural psychology (pp. 19 –29). Amsterdam, The Netherlands: Swets and Zeitlinger.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and goodness-of-fit testing in 2n contingency tables: A unified framework. Journal of the American Statistical Association, 100, 1009–1020. doi:10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. Psychometrika, 71,713–732. doi:10.1007/s11336-005-1295-9

McLeod, L. D., Swygert, K., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), Test scoring (pp. 189–216). Mahwah, NJ: Lawrence Erlbaum.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA & Oxford: Blackwell.

Michonski, J. D. (2011). Borderline personality disorder criteria in a population-based sample of 11 to 12-year-old children: An item-level analysis. Ph.D. dissertation, University of Houston, Houston, TX.

Miller, H. A. (2000). The development of the Miller's Forensic Assessment of Symptoms Test: A measure of malingering mental illness. *Dissertation Abstracts International: The Sciences and Engineering, 60, 4238- 4384.*

Miller, H. A. (2001). *MFAST: Miller Forensic Assessment of Symptoms Test professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

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

Mossman, D. (2003). Daubert, cognitive malingering, and test accuracy. *Law and Human Behavior*, *27*(3), 229-249.

Nijenhuis, E. R. S., Van der Hart, O., & Kruger, K. (2002). The psychometric characteristics of the Traumatic Experiences Checklist (TEC): First findings among psychiatric outpatients. *Clinical Psychology and Psychotherapy, 9, 200-210.*

Paek, I., Baek, S., & Wilson, M. (2012). An IRT modeling of change over time for repeated measures item response data using a random weights linear logistic test model approach. *Asia Pacific Education Review*, *13*(3), 487-494.

Rogers, R. (Ed.) (1997). Clinical assessment of malingering and deception. New York: Guilford Press

Sarason, I. G., Johnson, J. H., & Siegel, J. M. (1978). Assessing the impact of life changes: Development of the Life Experiences Survey. *Journal of Consulting and Clinical Psychology, 46, 932-946.*

Sharp, C., Michonski, J., Steinberg, L., Fowler, J. C., Frueh, B. C., & Oldham, J. M. (2014). An investigation of differential item functioning across gender of BPD criteria. Journal Of Abnormal Psychology, 123(1), 231-236. doi:10.1037/a0035637

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. Psychological Methods, 11, 402–415. doi:10.1037/1082-989X.11.4.402

Thissen, D. (2009). The MEDPRO project: An SBIR project for a comprehensive IRT and CAT software system—IRT software. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models (pp.67–113). In P. W. Holland & H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*(3), 291-307.

Vitacco, M. J., Rogers, R., Gabel, J., & Munizza, J. (2007). An evaluation of malingering screens with competency to stand trial patients: A known-groups comparison. *Law and Human Behavior, 31*(3), 249-260.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. Educational Measurement, 29, 39–47. doi:10.1111/j.1745-3992.2010.00182.x

Widiger, T., & Spitzer, R. (1991). Sex bias in the diagnosis of personality disorders. Clinical Psychology Review, 11, 1–22.