Examining the Potential for Gender Bias in the Prediction of Symptom Validity Test Failure by MMPI-2 Symptom Validity Scale Scores

Tayla T. C. Lee and John R. Graham
Kent State University
Tayla T. C. Lee, Department of Psychology, Kent Hall, Kent State University, Kent, OH 44242. E-mail: tconnerl@gmail.com

Martin Sellbom
The University of Alberta
Roger O. Gervais
Neurobehavioral Associates, Edmonton, Alberta, Canada, and University of Alberta

Using a sample of individuals undergoing medico-legal evaluations (690 men, 519 women), the present study extended past research on potential gender biases for scores of the Symptom Validity (FBS) scale of the Minnesota Multiphasic Personality Inventory–2 by examining score- and item-level differences between men and women and determining the extent to which FBS scores were able to correctly identify men and women who were divided into credible responders (n = 837) and noncredible responders (n = 372) on the basis of performance on symptom validity tests. Results indicated that women had slightly higher raw FBS scores than men (d = .29), and significant differences between men and women in item endorsement were demonstrated for 14 FBS items. Step-down hierarchical logistic regression procedures indicated predictive bias (χ²Δ = 23.72, p < .001). Follow-up analyses indicated intercept bias (χ²Δ = 23.51, p < .001) but not slope bias (χ²Δ = .22, p = .64). However, using the test publisher’s recommended FBS cutoff scores (Ben-Porath, Graham, & Tellegen, 2009), classification accuracies were similar for women and men (T > 80, h = −.02; T > 100, h = −.22, respectively). On the basis of these results, we conclude there is no evidence of clinically meaningful bias in predictions of symptom validity test failure made using FBS scores for men and women.

Keywords: Minnesota Multiphasic Personality Inventory–2, Symptom Validity scale (FBS), symptom validity tests, gender bias

The Symptom Validity scale (FBS) of the Minnesota Multiphasic Personality Inventory–2 (MMPI-2; Butcher et al., 2001) was developed by Lees-Haley, English, and Glenn (1991) to detect noncredible reporting of emotional difficulties by personal injury and other disability claimants. Created using a combination of rational, content-based item selection and endorsement rates for individuals thought to be malingering versus those believed to be responding honestly, the final scale consisted of 43 items, including statements assessing somatic and cognitive complaints as well as positive self-presentation (e.g., disavowal of a cynical world view, antisocial behavior, and substance use).

Although the original purpose of FBS was to identify exaggerated or feigned emotional distress and difficulties among personal injury claimants (Lees-Haley et al., 1991), subsequent research has suggested scores on this scale are not effective for this purpose (Ben-Porath, Graham, & Tellegen, 2009; Bury & Bagby, 2002; Efendov, Sellbom, & Bagby, 2008; Rogers, Sewell, Martin, & Vitacco, 2003). However, FBS scores have been demonstrated to be helpful in assessing noncredible reporting of cognitive and somatic difficulties (Ben-Porath, Graham, & Tellegen, 2009; Geffenstein, Baker, Gola, Donders, & Miller, 2002). In fact, FBS scores have been demonstrated in large meta-analyses to best predict exaggerated or feigned cognitive and somatic symptoms, when compared with the prediction of these types of noncredible responding by other MMPI-2 Validity scale scores (e.g., Infrequency–Psychopathology; Nelson, Hoelze, Sweet, Arbisi, & Demakis, 2010; Nelson, Sweet, & Demakis, 2006).

There have been many criticisms of FBS, including lack of construct validity, false identification of medical patients as noncredible responders, and overidentification of women as noncredible responders when compared with men (see Butcher, Gass, Cumella, Kally, & Williams, 2008, for a summary). Concerning gender bias, criticisms largely appear to originate from the scale developers’ recommendations for the use of different raw cut scores as indicative of noncredible responding for men and women (Williams, Butcher, Gass, Cumella, & Kally, 2009). Several authors have examined scale score and item endorsement rates between men and women in settings where the MMPI-2 is regularly used (Butcher et al., 2008; Nichols, Greene, & Williams, 2009; Williams et al., 2009). Overall, results of these studies indicated...
women typically had slightly higher raw scores on the FBS scale when compared with men. Women were also more likely to endorse items with content related to the disavowal of cynicism, substance use, and antisocial practices as well as items affirming some somatic complaints. However, these studies were limited as the authors did not have external criterion measures that would have permitted comparison of classification accuracies for men versus women.

Score and item endorsement differences do not necessarily indicate differences in predictions made using a scale’s scores (Greene, 1987; Prichard & Rosenblatt, 1980), and only one previous study has examined potential gender biases in predictions made with FBS scores. Presented in a reply to criticisms and using a known-groups design, Ben-Porath, Greve, Bianchini, and Kaufman (2009) examined the question of predictive differences for men and women in two large samples where FBS is frequently used. In their sample of individuals evaluated for difficulties related to mild traumatic brain injuries, the authors demonstrated significantly increased sensitivity of FBS scores in detecting cognitive/somatic overreporting for women compared with men, suggesting FBS scores were more accurate in correctly identifying noncredible responding in women. They demonstrated no differences in sensitivity of FBS scores in detection of cognitive/somatic malingering between genders in a sample of individuals undergoing evaluation for chronic pain. The authors suggested that, overall, these results did not support claims that FBS scores were biased against women.

The purpose of the present study was to extend research on potential gender differences in FBS scores into a non-neuropsychological disability evaluation setting. Using symptom validity tests (SVTs) to establish criterion groups of probable credible and noncredible responders,1 we first examined the effect of gender on FBS scores, comparing FBS scores and item endorsement rates. We then examined the accuracy of FBS scores in predicting SVT failure for men versus women. On the basis of previous research, we hypothesized that women would have higher raw scores on FBS than men. We also expected that women would endorse items related to cynicism, substance use, and antisocial practices with less frequency than men and that they would endorse items related to somatic complaints more frequently than men. Lastly, in congruence with previous research (Ben-Porath, Greve, et al., 2009), we hypothesized that gender would statistically moderate the relationship between FBS scores and credible versus noncredible response group membership (as defined by the absence or presence of SVT failure), but we did not expect this moderation to have a large effect size.

Method

Participants

Participants were selected for inclusion in the current study from a large archival database of 3,219 consecutive disability claimants referred for medico-legal psychological assessment at an independent practice in Edmonton, Alberta, Canada. Each individual provided informed consent to the psychological assessment and for their information to be used for research purposes.2 In keeping with ethical guidelines for the use of archival data, the final data set contained no identifiable personal information, and the study was approved by the institutional review board at Kent State University. Subsets of this data set have been used in previous studies on SVT performance in chronic pain patients (Gervais, Green, Allen, & Iversen, 2001; Gervais, Rohling, Green, & Ford, 2004), development and validation of the Response Bias scale for the MMPI-2-Restructured Form (Gervais, Ben-Porath, Wygant, & Green, 2007, 2008; Gervais, Ben-Porath, Wygant, & Sellbom, 2010), validation of the MMPI-2-Restructured Form (Gervais, Ben-Porath, & Wygant, 2009; Gervais, Wygant, Sellbom, & Ben-Porath, 2011), and posttraumatic stress disorder (Demakis, Gervais, & Rohling, 2008). In addition, Greiffenstein, Baker, Axellrod, Peck, and Gervais (2004) used these data to examine the association between MMPI-2 Validity scales (including FBS) and SVT failure, and descriptive information on FBS scores in this sample was provided in the FBS monograph (Ben-Porath, Graham, & Tellegen, 2009).

From the larger data set, we extracted a subsample of individuals who had completed all of the instruments used in the current study, including the MMPI-2 and each of the three SVTs described later. This procedure resulted in the inclusion of 1,278 consecutive cases assessed between January 1999 and August 2009. In keeping with recommended MMPI-2 interpretive guidelines, participants were excluded on the basis of inconsistent responding as detected by MMPI-2 Validity scales (i.e., Cannot Say [CNS] ≥ 30, Variable Response Inconsistency [VRIN] T ≥ 80, and/or True Response Inconsistency [TRIN] T ≥ 80; Butcher et al., 2001). However, because of the nature of the current study, participants were not excluded for content-based response styles. Using these criteria, 69 (5.4%) of the participants were excluded. Subsequent analyses indicated there were no significant differences between excluded and included participants in terms of gender, age, or years of education.

The final sample consisted of 690 men and 519 women referred for worker’s compensation (70.5%) and other medico-legal (29.5%) evaluations. None of these individuals were evaluated in the context of a criminal proceeding or to determine disability related to head injury. Included participants ranged in age from 17 to 73 years (M = 40.43, SD = 11.02) and reported having an average of 11.96 (SD = 2.54) years of education. Worker’s compensation referrals were seen for psychological assessment to assist in determining eligibility for or maintenance of compensation or other disability benefits and services related to psychological conditions arising from workplace injuries (musculoskeletal and/or orthopedic injuries, motor vehicle collisions, and/or other incidents, such as robberies, assaults, and workplace conflicts).

1 As discussed by Rogers (2008), noncredible responding is not a unitary construct and can include exaggerated or false reports of any combination of emotional, somatic, and cognitive symptoms as well as feigned or exaggerated reports of distress and impairment. In the current study, our use of the term noncredible is limited to the noncredible report of cognitive and somatic complaints as well as negative response bias, as detected by SVTs. As such, our method of group assignment does not preclude that individuals assigned to the credible responders group could have been exaggerating or fabricating purely psychological symptoms during the evaluation.

2 It is unknown how many claimants declined consent for their data to be incorporated into the archival data set, as no records were retained for these cases.
Fewer than 5% of female claimants were presenting with claims related to sexual harassment. Medico-legal referrals involved personal injury claims for psychological damages arising from motor vehicle collisions, with approximately 70% of assessments requested through plaintiff attorneys. Diagnoses were based on criteria from the Diagnostic and Statistical Manual of Mental Disorders (4th ed., American Psychiatric Association, 1994, and 4th ed., text revision, American Psychiatric Association, 2000), following an extensive clinical interview, review of psychological test data (which included the MMPI-2), and accompanying medical or other third-party documentation. The most frequently identified primary diagnoses were pain disorder (32.2%), anxiety-related disorders (e.g., adjustment disorder with anxiety or posttraumatic stress disorder, 39.4%), and mood disorders (e.g., major depressive disorder or adjustment disorder with depressed mood, 19.3%).

Measures

MMPI-2 (Butcher et al., 2001). The MMPI-2 is a 567 item, true/false, self-report inventory assessing social, behavioral, and emotional functioning. The instrument is one of the most widely used measures of personality and psychopathology in clinical practice (Camara, Nathan, & Puente, 2000). The current study included only the FBS of the MMPI-2, which is discussed later.

FBS. The FBS was introduced in 1991 by Lees-Haley et al. and was adopted as part of the standard scoring of MMPI-2 scales in 2006. Previous research has demonstrated FBS scores are internally consistent and are a significant predictor of noncredible cognitive and physical symptom reports (Ben-Porath, Graham, & Tellegen, 2009; Nelson et al., 2006, 2010). In the current sample, internal consistency estimates for FBS scores were acceptable, with \( \alpha = .76 \) (men) and \( \alpha = .72 \) (women).

Word Memory Test (WMT; Green, 2003; Green, Allen, & Astner, 1996; Green & Astner, 1995). The WMT is a computer-administered verbal memory SVT intended to assist an examiner in discriminating between individuals with genuine memory problems and those exhibiting incomplete effort or response bias associated with feigned memory deficits. The test has demonstrated high levels of sensitivity and specificity for response bias in simulator studies (e.g., Brockhaus & Merten, 2004; Tan, Slick, Strauss, & Hultsch, 2002). The administration and scoring software contains recommended scores for determining response bias (Green, 2003). These scores were used to evaluate the response validity of the respondents’ WMT scores in the current study.

Computerized Assessment of Response Bias (CARB; Allen, Conder, Green, & Cox, 1997). The CARB is an SVT that uses a forced choice digit recognition task to detect insufficient effort and response bias associated with exaggeration of memory problems. The test has good sensitivity (74%) and specificity (91%) in predicting WMT failure (Allen et al., 1997). The cutoff scores for biased responding specified in the test’s manual were utilized in this study.

Test of Memory and Malingering (TOMM; Tombaugh, 1996). The TOMM is an SVT using a forced choice visual recognition task designed to assist in discrimination between individuals with genuine and feigned memory impairment. We used the cutoff scores recommended in the test’s manual to identify biased responding in the current study. Previous research has demonstrated that the TOMM is sensitive to insufficient effort and negative response bias in a variety of populations, including individuals with presenting problems related to traumatic brain injury, toxic exposure, chronic pain, and psychiatric difficulties in forensic settings (Gierok, Dickson, & Cole, 2005; Greve et al., 2006; Haber & Fichtenberg, 2006; Iverson, Le Page, Koehler, Shojaian, & Badii, 2007).

Procedures

As part of a comprehensive 2-day evaluation, individuals were administered a psychological assessment battery consisting of the MMPI-2, several SVTs (including the WMT, CARB, and TOMM), and a variety of other cognitive, somatic, and psychopathological symptom self-report measures, depending on the focus of the assessment. Individuals referred for medical disability, workers’ compensation, or medical–legal purposes completed a core psychological assessment battery as described in Gervais et al. (2009). Individuals referred for chronic pain assessments completed the same core battery of tests, though fewer cognitive tests were administered. Additionally, modifications were made to the test battery in cases of noncompliance, specific physical disabilities, and educational or linguistic limitations.

We determined that substantial external incentives were present for the entire sample. All individuals in this study were seen in the context of active disability claims or personal injury litigation and were either receiving or pursuing disability benefits or awards. The assessment was typically requested to address questions related to validity of symptom presentation, psychiatric diagnosis, and necessary work restrictions or modifications as well as their relationship to the reportedly disabling accident or event and the presence of pre-existing conditions. In cases already receiving temporary total disability benefits, a finding of fitness for return to either preaccident or modified duties could result in a change of claim status, including return to work protocols and a systematic reduction and termination of benefits. For individuals whose claim status had not yet been determined, results of the assessment could contribute to adjudicative decisions to deny benefits or services. In view of these contingencies, incentives for individuals to either maintain existing disability benefits or present in a manner that would support a favorable decision on pending claims could not be ruled out. Similar approaches to the definition of external incentives have been used by other researchers (e.g., Bianchini, Greve, & Glynn, 2005; Greve, Ord, Bianchini, & Curtis, 2009; Slick, Sherman, & Iverson, 1999).

For the purposes of this study, individuals were assigned to one of two response style groups, credible or noncredible responders. Specifically, individuals were assigned to the noncredible responder group if they scored below chance or below the recommended cut score or scores on one or more of the SVTs (e.g., WMT, CARB, or TOMM). Although SVT failure can occur in any context and may or may not be necessarily related to the noncredible report of purely psychiatric symptoms, this method of identification has been suggested for identifying exaggerated or feigned ability deficits (Bianchini, Greve, & Glynn, 2005; Heilbroner et al., 2009; Slick et al., 1999). Further, Wygant et al. (2007) demonstrated that SVT failure was associated with overreporting of somatic and cognitive symptoms but not psychiatric symptoms in...
personal injury and disability claimants. As such, this method of identifying noncredible responders, although having limited utility for detecting purely overreported psychological symptoms, is appropriate for use in the current study, as all the included individuals were claiming significant impairment in their abilities to perform daily routines and work-related duties. For example, individuals in this sample scored, on average, much higher on scales assessing both general and work-impairing memory problems on the Memory Complains Inventory (Green, 2004) than individuals with medically verified, severe traumatic brain injuries.

Given that substantial external incentives were present for all individuals in the sample, using SVT performance as a criterion for noncredible response group assignment allowed the identification of definite (in the case of individuals who scored below chance levels on an SVT) or probable (in the case of individuals scoring below a particular recommended cut score on an SVT) response bias using the criteria for malingered neurocognitive dysfunction outlined by Slick et al. (1999). These criteria have also been suggested by Bianchini et al. (2005) for the identification of malingered pain-related disability as well as by the American Academy of Clinical Neuropsychology for screening of noncredible reporting of cognitive symptoms and ability deficits (e.g., difficulties with attention and memory) in individuals reporting depressive and/or anxiety related disabilities (Heilbroner et al., 2009). Further, this method of group assignment is congruent with recommendations issued within the consensus statement on SVT use issued by the National Association of Neuropsychology (Bush et al., 2005). Bush et al. (2005) asserted that SVTs were forensically and medically necessary in all evaluations of disability, because they assess not only cognitive effort and noncredible cognitive symptoms but also because poor performance on SVTs is associated with unreliable self-reports. As such, they concluded that the possibility of deliberate response distortion (i.e., a negative response bias) should be considered when an SVT score falls below recommended levels, no matter the presenting problem. Lastly, this method of group assignment is strongly supported by evidence of a nonspecific association between SVT failure and scores on MMPI-2 indices of overreported somatic complaints, depressive symptomatology, and general emotional distress in individuals undergoing evaluation for disability as an alleged result of neurological (e.g., traumatic brain injury), medical (e.g., pain), or psychological (e.g., depression or posttraumatic stress disorder) difficulties (Nelson, Sweet, Berry, Bryant, & Granacher, 2007; Thomas & Youngjohn, 2009).

This method of response group assignment resulted in identification of 372 probable noncredible responders and 837 credible responders. Of the claimants in the sample assigned to the noncredible responders group, 19 scored below chance on one or more of the SVTs. Of the remaining 353 individuals in the noncredible responders group, 217 scored below recommended cut scores on one SVT, whereas 136 scored below the cutoffs on more than one SVT. The base rate of noncredible responding (30.8%) in this sample was comparable with previously published rates of noncredible responding in medicolegal evaluations (Mittenberg, Patton, Canyock, & Condit, 2002). Women were less likely to be identified as noncredible responders using this procedure, though the difference was small in effect size, $\chi^2(1, N = 1209) = 11.29, p < .001$, odds ratio [OR] = 0.65. Table 1 provides a breakdown of the credible and noncredible responder groups’ demographic and diagnostic characteristics (which were assigned using all available test information, including SVT and MMPI-2 results) as well as results of analyses examining potential differences in these characteristics for the combined sample and by gender.

**Results**

**Score- and Item-Level Differences.** To examine whether men and women differed on mean raw FBS scale scores, we conducted two t tests and then used Cohen’s d (1988) to quantify the size of obtained effects (.3 = small effect, .5 = medium effect, .8 = large effect). Results of the first t test examining FBS raw score differences for men and women when both credible and noncredible responders were included in the sample indicated a small, statistically significant difference, $t(1, 1207) = -5.26, p < .001, d = -.29$. On average, men ($M = 23.25, SD = 6.01$) scored lower than women ($M = 25.01, SD = 5.40$). These raw scores correspond to T scores equal to 80 for men and 77 for women.

As potential differences could have been due to inflation of mean scores by the inclusion of noncredible responders in the analysis, we also examined potential gender differences in FBS raw scores for only the credible responders. Results were statistically significant with a small effect size, $t(1, 835) = -5.98, p \leq .001, d = -.41$. When only credible responders were considered, men ($M = 21.93, SD = 5.92$) continued to score lower than women ($M = 24.27, SD = 5.31$) on average. These raw scores correspond to T scores of 77 and 75 for men and women, respectively.

Item-level differences. We next examined potential FBS item endorsement differences between men and women using a series of chi-square analyses, first calculating these analyses with the entire sample and then again with only the credible responder group. The practical significance of the chi-square results was interpreted by calculating ORs and by examining the 95% confidence intervals for the ORs. Because of the number of chi-square tests calculated, a Bonferroni correction was applied to correct for potential Type I error, and the required level for statistical significance was $p < .001$ (.05/43).

Results of the statistically significant chi-square analyses examining FBS item endorsement differences by gender are presented in Table 2. Overall, results indicated that men were more likely than women to endorse items in the keyed direction related to coughing blood, escapism, and suicidal thoughts, with ORs ranging from 0.57 to 1.74. Women were more likely to than men to endorse items in the keyed direction related to having many headaches, being easily tired, and having poor energy as well as items disavowing cynicism, antisocial beliefs, and alcohol use. ORs for these items ranged from 1.63 to 3.21. The pattern of

---

3 We are not suggesting that individuals claiming disability because of psychiatric syndromes should be assessed solely with SVTs to screen for noncredible responding. Rather, we are arguing that SVTs provide one source of information regarding the noncredible report of ability deficits that should be obtained during a comprehensive assessment performed for purposes of a disability or personal injury evaluation.
622 LEE, GRAHAM, SELLBOM, AND GERVAIS

Demographic Characteristics of the Study Sample by Credible (0 SVT Failures) and Noncredible (≥1 SVT Failures) Responders

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Credible responders</th>
<th>Noncredible responders</th>
<th>p</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N = 837)</td>
<td>(N = 256)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined genders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (M)</td>
<td>39.76 (10.98)</td>
<td>41.93 (11.00)</td>
<td>.002</td>
<td>.27</td>
</tr>
<tr>
<td>Age (SD)</td>
<td>10.98</td>
<td>11.00</td>
<td>.10</td>
<td>.13</td>
</tr>
<tr>
<td>Education (M)</td>
<td>12.07 (2.50)</td>
<td>11.68 (2.61)</td>
<td>.01</td>
<td>.15</td>
</tr>
<tr>
<td>Education (SD)</td>
<td>2.50</td>
<td>2.61</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>Diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood</td>
<td>12.07 (2.50)</td>
<td>11.68 (2.61)</td>
<td>.01</td>
<td>.15</td>
</tr>
<tr>
<td>Mood (SD)</td>
<td>2.50</td>
<td>2.61</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>Anxiety</td>
<td>2.50</td>
<td>2.61</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>Other</td>
<td>2.50</td>
<td>2.61</td>
<td>.11</td>
<td>.12</td>
</tr>
</tbody>
</table>
| Note: For age and education, t-tests and Cohen’s d effect size are reported. For diagnosis, chi-square analysis is reported. Diagnoses are intended for descriptive purposes only, as they were based on available test information, including measures in the current study. SVT = symptom validity test; ES = effect size.

Results of the step-down hierarchical logistic regression analyses are presented in Table 3. The first step in the process consisted of testing for evidence of predictive differences. The test of prediction differences compared the prediction of group membership by FBS scores alone with the full model containing FBS scores, gender, and the interaction of FBS scores and gender. As seen in Table 3, comparison of the reduced model with the full model (i.e., the Prediction $\chi^2$) was statistically significant. However, as indicated by the beta weights, only FBS scores and gender were significant individual predictors of credible/noncredible responder group membership. Higher FBS scores, as well as being female, were related to a higher probability of being classified as a noncredible responder compared with being classified as a credible responder, with ORs (i.e., $Exp[B]$) of 1.98 and 1.87 for FBS scores and gender, respectively.

Additional models were then tested to determine if the significant result was due to slope and/or intercept differences. The analysis examining slope differences, the second step in the process, compared a model with MMPI-2 FBS scores and gender with the model containing FBS scores, gender, and the interaction of FBS scores and gender in the prediction of response group membership. Statistically significant results supporting discrepant slope values would indicate FBS scores were differentially associated with prediction of response group membership for men and women. As seen in Table 3, in the prediction of credible/noncredible responder group membership results provided no support for statistically significant differences in slope values between men and women.

Because there was no evidence of slope differences, the last step of the process was an examination of potential intercept differences in which the prediction of response group membership by FBS scores alone was compared with results from the model containing FBS scores and gender. Results supporting discrepant intercept values would indicate FBS scores consistently under- or overpredicted group membership when applied to the different genders. As seen in Table 3, results indicated the addition of gender to the prediction of group membership to the model containing FBS scores was statistically significant and supported a discrepancy in intercept values for men and women. Inspection of logit weights ($B$) and associated ORs ($Exp[B]$) in this model suggested a greater likelihood of being classified as a noncredible

---

4 These results are available from the first author.
Table 2

Statistically Significant Item Endorsement Frequencies/Differences for FBS by Gender for Entire Sample

(Men = 690, Women = 518)

<table>
<thead>
<tr>
<th>MMPI-2 Item</th>
<th>% of men</th>
<th>% of women</th>
<th>chi2</th>
<th>Odds ratio</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>58 (F), not cynical</td>
<td>45.3</td>
<td>63.2</td>
<td>37.63*</td>
<td>2.07</td>
<td>1.64</td>
</tr>
<tr>
<td>81 (F), not cynical</td>
<td>37.8</td>
<td>57.8</td>
<td>47.53*</td>
<td>2.25</td>
<td>1.79</td>
</tr>
<tr>
<td>110 (F), not cynical</td>
<td>35.0</td>
<td>47.2</td>
<td>18.17</td>
<td>1.66</td>
<td>1.31</td>
</tr>
<tr>
<td>117 (F), cough blood</td>
<td>28.9</td>
<td>18.9</td>
<td>16.08</td>
<td>0.57</td>
<td>0.43</td>
</tr>
<tr>
<td>152 (F), easily tired</td>
<td>77.5</td>
<td>87.0</td>
<td>17.79</td>
<td>1.95</td>
<td>1.42</td>
</tr>
<tr>
<td>176 (F), many headaches</td>
<td>60.8</td>
<td>73.4</td>
<td>21.17</td>
<td>1.78</td>
<td>1.39</td>
</tr>
<tr>
<td>250 (F), dislike criminals</td>
<td>71.3</td>
<td>88.8</td>
<td>54.67*</td>
<td>3.21</td>
<td>2.33</td>
</tr>
<tr>
<td>264 (F), minimal alcohol</td>
<td>64.2</td>
<td>82.6</td>
<td>49.92*</td>
<td>2.65</td>
<td>2.01</td>
</tr>
<tr>
<td>284 (F), truthful</td>
<td>46.4</td>
<td>51.5</td>
<td>16.84*</td>
<td>1.62</td>
<td>1.28</td>
</tr>
<tr>
<td>374 (F), not cynical</td>
<td>40.8</td>
<td>55.1</td>
<td>24.39*</td>
<td>1.78</td>
<td>1.42</td>
</tr>
<tr>
<td>419 (F), not vengeful</td>
<td>49.1</td>
<td>63.6</td>
<td>25.04*</td>
<td>1.81</td>
<td>1.43</td>
</tr>
<tr>
<td>505 (T), escape problems</td>
<td>39.0</td>
<td>27.0</td>
<td>19.13</td>
<td>1.73</td>
<td>1.35</td>
</tr>
<tr>
<td>506 (T), suicidal thinking</td>
<td>27.4</td>
<td>17.4</td>
<td>16.62*</td>
<td>1.79</td>
<td>1.35</td>
</tr>
<tr>
<td>561 (F), poor energy</td>
<td>55.5</td>
<td>67.1</td>
<td>16.51*</td>
<td>1.63</td>
<td>1.29</td>
</tr>
</tbody>
</table>

Note. Percentage columns represent the number of individuals within each gender answering in the keyed direction. MMPI-2 = Minnesota Multiphasic Personality Inventory–2; FBS = Symptom Validity scale of the MMPI-2; T = true; F = false.

*p < .001.

responder compared with a credible responder as FBS scores increased (B = .66, p < .001, Exp[B] = 1.93) and suggested that the odds of being classified as a noncredible responder were significantly greater for women compared with men (B = .65, p < .001, Exp[B] = 1.91).

Classification accuracies and comparisons. The regression analyses described earlier allowed an examination of the prediction of group membership but did not allow for the quantification of the accuracy of those predictions. To determine the practical significance of predictive differences (i.e., the accuracy of the group membership predictions), using the formulas outlined by Meehl and Rosen (1955) we calculated classification accuracies, including overall correct classification, sensitivity, and specificity as well as negative and positive predictive powers, for the prediction of response group using the test publisher’s recommended FBS score cutoffs of T > 80 and T > 100 (Ben-Porath, Graham, & Tellegen, 2009).

However, our interest in the current study was in contrasting classification accuracies between the genders to determine the impact of any predictive biases demonstrated in previous regression analyses, not in inspecting the classification accuracies themselves. As such, we compared the obtained accuracy of classification statistics for men and women using Cohen’s h statistic (Cohen, 1988). Cohen’s h is an effect size statistic reflecting the magnitude of the difference between two proportions, with values greater than .2/.3, .5, and .8 reflecting small, medium, and large differences, respectively.

Calculated classification accuracy statistics and comparisons of these proportions for men and women are presented in Table 4. Overall, results indicated that at T > 80, there were negligible-to-small differences in classification accuracies between men and women (h = -.05 to .26), with no meaningful differences in overall correction classification (h = -.02). Further, at a T > 100, results indicated negligible-to-small differences in the classification accuracies (h = -.29 to .32), with only small differences in overall correct classification (h = -.20). This result indicated more men than women were correctly classified into response style groups, although women who were, in fact, noncredible responders

Table 3

Logistic Regression Analyses of Credible (0 SVT Failures) Versus Noncredible (>1 SVT Failures) Responder Group Status on Minnesota Multiphasic Personality Inventory–2 (MMPI-2) Symptom Validity (FBS) Scale, Gender, and FBS × Gender Interaction Term (N = 1,209)

<table>
<thead>
<tr>
<th>FBS</th>
<th>Gender</th>
<th>FBS × Gender</th>
<th>chi2</th>
<th>Prediction</th>
<th>Slope</th>
<th>Intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>.68*</td>
<td>.63*</td>
<td>-.07</td>
<td>107.35*</td>
<td>23.72*</td>
<td>.22</td>
<td>23.51*</td>
</tr>
</tbody>
</table>

Note. FBS was standardized prior to analyses using z scores. Gender is coded 0 for men and 1 for women. Regression weights are displayed in the three columns labeled B. Gender = reported sex; FBS × Gender = interaction term; \( \chi^2 \Delta \) = the change in the model \( \chi^2 \) accounted for by the addition(s) to the model. SVT = symptom validity test. *p < .001.
were more likely to have FBS scores below the recommended cut points ($h$ for specificity at $T > 80$ and $T > 100$ was $-.16$ and $-.29$, respectively).

**Discussion**

Overall, results of this study indicate FBS scores are not differentially accurate in identifying SVT failure in men and women. Women had slightly higher raw FBS scores than men, though these differences were negligible when converted to MMPI-2 $T$ scores. Significant differences in item endorsement between men and women were found for 14 FBS items, indicating women were more likely to endorse items in the keyed direction related to headaches, problems with energy/motivation, and a disinvolvement of cynicism, antisociality, and alcohol use. These results are similar to those demonstrated in past research where differences were demonstrated in raw scale scores and item endorsements in various settings (Nichols et al., 2009; Williams et al., 2009).

As mentioned previously, scale score and item endorsement differences between men and women do not necessarily indicate that use of that scale’s scores will lead to predictive differences. When predictive differences were examined using regression analyses, results of the current study indicated that, statistically, gender was a moderator in the prediction of SVT performance. However, although intercept differences were demonstrated, utilizing the test publisher’s recommended FBS cutoff scores ($T > 80$ and $T > 100$; Ben-Porath, Graham, & Tellegen, 2009), results indicated classification accuracies were similar for women and men, as differences in classification were negligible-to-small in their effect. These results appear to parallel those demonstrated in previous research examining the effect of ethnicity on predictions made using MMPI-2 scales (e.g., Arbisi, Ben-Porath, & McNulty, 2002) where, although predictive differences were demonstrated, they had little to no practical meaning. This pattern of results is also in congruence with Ben-Porath, Greve, et al.’s (2009) demonstration of no practically meaningful differences in classification between men and women in a somatic pain sample. Lastly, a close examination of the pattern of classification differences in the current study would suggest some support for a finding similar to the traumatic brain injury group included in Ben-Porath, Greve, et al.’s (2009) study, as specificity was greater for women, indicating that more women than men who failed one or more SVT scored below $T > 80$ and $T > 100$, though differences between specificity statistics for men and women were of negligible-to-small effect sizes.

We believe it is important to note that in the current study, mean FBS $T$ scores for credible responders (i.e., those not scoring below chance or recommended cut scores on any included SVT) were high ($T > 77$ and 75 for men and women, respectively), as by themselves these scores have an important implication. The high mean FBS $T$-score elevations for both men and women in the credible responders group are directly reflected in the demonstrated low sensitivity and specificity values in the classification results. However, the relatively high scores in honest responders in this sample are not completely unexpected, likely having been influenced by legitimately experienced symptoms as well as by the presence of external incentives. Previous research has demonstrated that FBS scores tend to be higher in individuals with medically related complaints (see, e.g., Tables 2–4 in Ben-Porath, Graham, & Tellegen, 2009) as well as directly and positively related to the presence of external incentives (Ben-Porath, Greve, et al., 2009). Further, though the relatively high FBS scores in the current study lowered the obtained classification values, these rates are not lower than would be expected given the base rate of noncredible responding defined in this sample. For example, these rates are comparable with those obtained in published, known-groups studies of individuals with pain-related symptoms (e.g., Bianchini, Etherton, Greve, Heinly, & Meyers, 2008). Nonetheless, the low sensitivity and specificity of FBS scores demonstrated in the current study directly support Ben-Porath, Graham, and Tellegen’s (2009) caution regarding over-interpreting FBS scores and against using scores on this scale in isolation to determine the credibility of somatic and cognitive symptom reports.
In addition to concerns with the version of FBS scored on the MMPI-2, Butcher et al. (2008) have asserted that the purported gender bias of FBS scores would carry over to the Symptom Validity Scale–Revised (FBS-r), which is scored on the Restructured Form of the MMPI-2 (Ben-Porath & Tellegen, 2008; Tellegen & Ben-Porath, 2008). Although not presented in the current article, we conducted the same analyses described for FBS using scores on FBS-r. Overall, results of these analyses conformed to the same pattern of results just discussed for FBS.5 There were raw score and item endorsement differences for men and women as well as statistical support for differences in predicting whether an individual had failed an SVT due to intercept bias. However, when classification accuracies were examined, the practical meaning of the demonstrated intercept difference was negligible to small, counter to the idea that scores on FBS-r are biased against women.

One limitation of the current study was that the sample used consisted of disability claimants who did not have head injuries. Using scores on FBS as a measure of noncredible cognitive symptom reports is based largely on the neuropsychological research with brain injury claims. In that context, FBS scores may well function differently. However, as reviewed by Ben-Porath, Graham, and Tellegen (2009) and supported by the recent meta-analysis conducted by Nelson et al. (2010), FBS scores are also sensitive to noncredible reporting of somatic symptoms in disability settings, which suggests that the results of the current study with claimants reporting disability not related to a head injury are likely generalizable as well.

Another consideration is the nature and number of SVTs used to assign response group membership. Other studies examining the validity of FBS scores have used alternative measures (e.g., Bianchini et al., 2008), which might correlate more strongly or be differentially predicted by FBS scores for the two genders. Because of sample size restrictions, we were only able to classify individuals as noncredible responders independently of the MMPI-2 using Slick et al.’s (1999) definite malingered neurocognitive classification for cases scoring below chance on any SVT and using probable response bias classification (Criterion B2) for individuals who failed one or more SVTs at established cutoffs and who, by definition, met the criteria for the presence of a substantial external incentive (Criterion A). It can be argued that different results might have been obtained if more stringent response group criteria had been used (e.g., failure of more than one SVT), as previous research has suggested use of multiple SVTs increases the identification of true positives while decreasing the identification of false positives (Greve, Ord, Curtis, Bianchini, & Brennan, 2008). However, it can also be argued that requiring failure on multiple SVTs is certainly appropriate when utilizing SVTs that have relatively low specificity but is unnecessarily restrictive when the SVTs in question already have an established high level of specificity in diverse clinical populations with significant cognitive impairment. Nonetheless, given the potential costs of misclassifying individuals as noncredible responders when they are, in reality, responding credibly, we suggest that future research continue to examine potential gender differences of FBS score predications in alternative samples and using various methods of defining noncredible group membership.

In summary, given that men and women have been demonstrated to have FBS raw score and item endorsement differences, some authors have called into question the utility of FBS (e.g., Williams et al., 2009). However, scale and item endorsement differences do not indicate whether a scale’s scores are biased against some group of individuals in the prediction of external criteria (Pritchard & Rosenblatt, 1980). On the basis of the results of the current and past studies examining predictive bias, we conclude there is no evidence for meaningful differences in the prediction of potential noncredible responders for men and women when using FBS scores as a predictor. Specifically, the current study demonstrated minimal differences in predictions of SVT failure in a sample of non-head-injury disability claimants and personal injury litigants between genders. Further, one previous study (Ben-Porath, Greve, et al., 2009) demonstrated minimal-to-small differences in the classification of men and women into malingers using formal classification systems for malingered neurocognitive dysfunction and malingered pain-related disability. Further, when biased predictions have been demonstrated, results have suggested that FBS scores actually underpredict credible response group membership to a greater extent for men compared with women—countering claims that FBS scores are biased against women. Overall, it appears that use of FBS scores recommended in the FBS test monograph (T > 80 and T > 100; Ben-Porath, Graham, & Tellegen, 2009) does not lead to differential prediction of noncredible group membership in medico-legal settings for men and women.

5 These results are available from the first author.

References


GENDER BIAS IN SYMPTOM VALIDITY TEST FAILURE

627


Received September 15, 2010
Revision received September 6, 2011
Accepted September 7, 2011