Validation of the MMPI-2 Response Bias Scale and Henry–Heilbronner Index in a U.S. Veteran Population

J. Christopher Young1,*, Lee A. Kearns2, Brad L. Roper1,*

1Memphis Veterans Affairs Medical Center, Memphis, TN, USA
2Western Washington Medical Group, Everett, WA, USA

*Corresponding author at: Memphis Veterans Affairs Medical Center, 1030 Jefferson Ave, Memphis, TN 38104, USA. Tel.: +1-901-523-8990; fax: +1-901-577-7371.
E-mail address: john.young5@va.gov (J.C. Young) or brad.roper@va.gov (B.L. Roper).
Accepted 19 February 2011

Abstract

The Response Bias Scale (RBS) and the Henry–Heilbronner Index (HHI) are two recently developed Minnesota Multiphasic Personality Inventory-2 (MMPI-2) validity scales empirically derived for the purpose of detecting feigned symptom report. Utilizing a veteran sample, the present study examined the utility of these and other MMPI-2 validity scales in predicting Word Memory Test (WMT) failure and presence of recent, current, or upcoming compensation evaluation. Although a significant predictor of WMT performance, RBS did not show incremental validity over Infrequency scale of the MMPI-2 in prediction of WMT failure and was not significantly associated with membership in the “Compensation-Context” (CC) group. In contrast, HHI best predicted CC group membership, but only trended toward significance in predicting WMT failure. In predicting CC group membership, HHI showed incremental validity above the established MMPI-2 validity scales, but its specificity was low. In the context of current literature on RBS and HHI, results support continued use of RBS and HHI but suggest that these scales may perform differently in samples other than the compensation-seeking samples on which they were developed.

Keywords: Minnesota Multiphasic Personality Inventory; Response bias; Military veterans; Neuropsychological assessment; Psychometrics

Introduction

Clinical neuropsychologists are held accountable for the validity of the data derived from evaluations and the confidence with which they can make determinations and recommendations from that data (Bush et al., 2005). Over the past decade, the profession has seen a dramatic rise in the amount of interest, published research, and measures dedicated to response bias. In contrast to clinical settings, which typically lack incentives for symptom exaggeration, forensic settings carry clear incentives to put forth misleading symptom presentations, resulting in a higher prevalence of response bias. As such, a preponderance of the research on response bias has focused on forensic settings (Larrabee, 2003; Mittenberg, Patton, Canyock, & Condit, 2002). These authors have estimated the base rate for malingering within forensic evaluations to be 30%–40%, and Iverson (2003) stated that “Any neuropsychological evaluation that does not include careful consideration of the patient’s motivation to give their best effort should be considered incomplete” (p. 138). Further, the American Academy of Clinical Neuropsychology (AACN) and National Academy of Neuropsychology (NAN) have issued policy papers recommending that employment of symptom validity tests (SVTs) become standard practice across assessment contexts (Bush et al., 2005; Heilbronner et al., 2009).

Loring, Lee, and Meador (2005) identified the lack of consensus among those practicing in general clinical settings regarding the use of SVTs. Patients in these evaluations often lack the clear incentives for secondary gain, and providers are under greater pressure than they are in medico-legal evaluations to minimize appointment duration due to limited assessment resources or constraints imposed by third-party payers. However, as Bush and colleagues (2005) noted, in general clinical
settings, an evaluation may have forensic relevance that is not apparent to the clinician but which can impact the likelihood of symptom fabrication.

Of particular significance to this study is the unique framework of the U.S. Department of Veterans Affairs (DVA), where a variety of manifest and latent influences may impact symptom presentation and potential response bias. The DVA includes the Veterans Health Administration (VHA) and the Veterans Benefits Administration (VBA), the latter of which grants some veterans a pension, similar to the benefits offered by Social Security, as well as disability compensation, which is a benefit paid because of injuries or diseases associated with active duty military service. With U.S. service members involved in extended conflicts in Iraq and Afghanistan, the VA has experienced a dramatic increase in the submission of compensation claims over the past decade, rising from 421,000 in 2000 to approximately 719,000 in 2008 (United States Government Accountability Office, 2010). For those who are deemed to have a 100% “service-connected” disability in 2010, the amount of this tax-free compensation exceeds $32,000 per year (United States Department of Veterans Affairs, 2009). Veterans may also receive substantial additional compensation for dependents and/or if deemed eligible for Special Monthly Compensation, which is awarded to individuals “disabled beyond a combined degree percentage or due to special circumstances (i.e., aid and attendance, loss of use of one hand, etc.)” (United States Department of Veterans Affairs, n.d.). Furthermore, service-connected veterans are eligible for increased access to medical care and social services depending on their level of service connection. After receiving disability compensation, veterans may later seek increases in their compensation based on a worsening of their condition, or they may be subject to reexamination to establish whether improvements in disabled functioning may result in a decrease in compensation (VA Compensation, n.d.). As such, many neuropsychological evaluations conducted within a general clinical framework of VA healthcare may be impacted by patient concerns regarding the attainment and/or maintenance of disability.

Attention has been given to the influence of compensation-seeking on the report of veterans applying for compensation due to Posttraumatic Stress Disorder (PTSD; Arbisi, Ben-Porath, & McNulty, 2006; Atkinson, Henderson, Sparr, & Deale, 1982; Frueh et al., 2003; Morel, 2008, 2010), but little has been published regarding factors related to neurocognitive effort in the VA’s unique evaluation setting. A number of researchers have recently suggested that compensation and pension (C&P) evaluations may impact report and presentation of neurocognitive symptoms in Operation Iraqi Freedom/Operation Enduring Freedom (OIF/OEF) veterans (Belanger, Uomoto, & Vanderploeg, 2009; Iverson, Langlois, McCrea, & Kelly, 2009). Additionally, findings of clear dissociations between over reporting of psychopathology and insufficient cognitive effort (Nelson, Sweet, Berry, Bryant, & Granacher, 2007) suggest that both should be routinely assessed during evaluations that may have implications for attainment or maintenance of disability.

The Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher et al., 2001) is widely used in both general clinical and forensic neuropsychological evaluations to measure psychopathology and emotional distress (Lees-Haley, 1992). The standard MMPI-2 validity scales also provide an indication of response styles, biases, and the over- or under-reporting of psychopathological symptoms (Butcher et al., 2001). However, the MMPI-2 is considered a poor measure of neurocognitive effort (Gervais, Wygant, & Ben-Porath, 2005). The Fake Bad Scale (FBS) on the MMPI-2 has been the subject of scrutiny regarding its ability to discriminate response bias from legitimate symptom report (Butcher, Gass, Cumella, Kally, and Williams, 2008; Williams, Butcher, Gass, Cumella, & Kally, 2009). The reader is referred to Ben-Porath, Greve, Bianchini, and Kaufmann (2009) and Ben-Porath, Greve, Bianchini, and Kaufmann (2010) for responses to such criticisms. Relevant to the current study, Gervais, Ben-Porath, Wygant, and Green (2007) empirically derived an MMPI-2 scale, the Response Bias Scale (RBS), by identifying 28 items from the MMPI-2 item pool that discriminated between individuals with passing or failing scores on three well-validated symptom validity measures, including the Word Memory Test (WMT; Green, 2003), the Test of Memory Malingering (TOMM; Tombaugh, 1996), and the Computerized Assessment of Response Bias (Allen, Conder, Green, & Cox, 1997). The sample for their scale development was composed of 1,212 consecutive non-head injury disability claimants and counseling clients referred to the first author’s private psychology practice. The authors subsequently validated RBS with a separate sample of 317 patients recruited from two private practices, one of which was also derived from the first author’s practice. Large proportions of both the development (76%) and validation (87%) samples were seeking compensation through litigation, workman’s compensation, or disability claims, and the samples were composed of individuals diagnosed a variety of psychiatric and medical (e.g., chronic pain, orthopedic) conditions. Internal consistency was 0.76 across both the development and validation samples (Gervais et al., 2007).

A number of subsequent studies have further validated RBS in the prediction of response bias and in discriminating between groups with and without secondary gain concerns. Nelson, Sweet, and Heilbronner (2007) found that RBS was successful in distinguishing between the secondary gain (civil litigants) and non-secondary gain to a greater extent than traditional MMPI-2 validity scales. Larrabee (2008) found that RBS best predicted performance relative to other MMPI-2 validity indices (i.e., FBS-RF, Henry–Heilbronner Index [HHI], Infrequent Somatic Response [Fs-RF], FBS, Infrequency [F], Infrequency-Psychopathology [Fp], Fptsd) on a series of embedded SVTs in samples of clinical patients and malingering.
More recently, Wygant and colleagues (in press) sought to validate the RBS in a criminal sample, noting that a comparison sample involved in disability litigation was also employed. Similar to prior studies, RBS displayed incremental validity in the prediction of SVT performance in civil litigants; however, it failed to display incremental validity in predicting SVT performance within the criminal sample. Nevertheless, RBS displayed strong specificity (0.89–0.91) at T score cutoffs >90 in the disability claimant group and at cutoffs >100 in the criminal sample. Sensitivity was adequate (0.38–0.59) for both groups. Large effect sizes were also found for RBS in predicting SVT failure in both groups (i.e., criminal, d = 1.48; disability, d = 1.24). As such, RBS has been validated for both civil and criminal litigation samples as well as discriminating between groups based on compensation-seeking status. To our knowledge, only a single study has investigated the validity of RBS within a Veterans Affairs Medical population. Whitney, Davis, Shepard, and Herman (2008) found that RBS outperformed existing F-family MMPI-2 scales in predicting SVT failure on the TOMM. However, the study did not address compensation-seeking as a variable itself within the VA population.

Butcher, Gass, Cumella, Kally, and Williams (2008) suggested that RBS is closely related to FBS and in their critique of the validity of FBS expressed concerns regarding the use of performance on a cognitive measure as an indicator of a pattern of responding on a personality measure. In doing so, they echoed the concerns expressed by Bush and colleagues (2005, p. 424) who stated that “Invalid performance on a measure of personality does not allow for an a priori conclusion that the neurocognitive test results are also unreliable, and vice versa.” This argument is not without merit, since traditional MMPI validity scales, such as the F scale, were not derived for the purpose of detecting symptom exaggeration. Nevertheless, subsequent research has found that some scales, such as F, have utility in detection of cognitive response bias (Ben-Porath, Greve, Bianchini, & Kaufman, 2009). Further, the statement by Bush and colleagues (2005) raises an empirical question whether response patterns on a personality measure, such as the MMPI-2, can display strong operational characteristics in the detection of response bias. As previously described, Gervais and colleagues (2007) sought to address limitations of the MMPI-2 by employing an empirical keying methodology in the construction of the RBS, whereby items were selected based upon their relationship with SVT performance. Although using SVT performance to select items may be of concern to Butcher and colleagues (2008), RBS was constructed using an empirically strong methodology, and its validity should be empirically investigated in a variety of populations and contexts.

The HHI (Henry, Heilbronner, Mittenberg, & Enders, 2006) is another recently developed, empirically derived scale purported to reflect a “pseudosomatic factor” (p. 786). The HHI is composed of 15 items selected from the FBS (Lees-Haley, English, & Glenn, 1991) and the Pseudoneurologic Scale (PNS; Shaw & Matthews, 1965) based upon the strength of correlations between items and classification of suspected malingering. Data for scale development were retrospectively gathered from 119 adults administered an MMPI-2 as part of a neuropsychological assessment, and individuals were categorized as non-malingering or “probable malingering” based upon SVT failure and meeting additional criteria specified by Slick, Sherman, and Iverson (1999) and Henry and colleagues (2006, p. 788). Henry and colleagues (2006) found HHI to be superior to both FBS and PNS in accurately classifying individuals to either the non-malingering or probable malingering groups, and it was reported to have strong specificity (0.89) and sensitivity (0.80) at the recommended cutoff. Henry, Heilbronner, Mittenberg, Enders, and Stanczak (2008) subsequently utilized HHI and FBS in conjunction with the Restructured Clinical Scale 1, Somatic Complaints (Tellegen et al., 2003), to examine predictive validity in the identification of individuals involved in personal injury or disability litigation. The HHI and FBS displayed relatively equivalent sensitivity and specificity. Resultantly, the authors suggest that the use of both scales may not be necessary given that the scales have many overlapping items and similar psychometric properties.

Whitney and colleagues (2008) examined the predictive relationship of MMPI-2 validity scales, including HHI, RBS, and FBS, on pass/fail performance on the TOMM in a veteran sample. It was found that RBS best predicted TOMM failure and displayed significant incremental validity when included with all other validity scales except HHI, where RBS still trended toward significance. Although HHI and Fb also displayed significant group differences between those who passed and failed the TOMM, they did not display incremental validity when added to a regression model already containing RBS. It is also noteworthy that HHI scores for the group that passed the TOMMs were above the recommended cutoff for symptom exaggeration (recommended cutoff = 8; mean HHI for those passing TOMM = 8.75; mean HHI for those failing TOMM = 11.2). Whitney and colleagues (2008) note that these findings are tempered by a number of limitations. First, the relatively small sample provided limited power, and this is particularly the case for the examination of alternative RBS cutoffs using receiver operating characteristic analysis. Second, the authors stated that subjects were consecutive referrals administered the TOMM and MMPI-2, and effort testing was utilized for reasons that included participants being under 65 years of age or having a history of requests for a C&P evaluation. Review of sample characteristics finds that all participants were under 65 years of age and the proportion involved in a C&P evaluation was not reported. As such, it appears that age was the primary selection criteria and the relationships among involvement in a compensation claim and variables of interests (e.g., TOMM and MMPI-2 validity scales) were not explored. The authors note that it is possible that the RBS may have identified...
individuals with a negative response bias who passed the TOMM given its relatively low sensitivity. If so, they suggest that the positive and negative predictive values of RBS are likely higher than they were able to demonstrate given the limitations of the sensitivity of the TOMM and suggest that future research with the RBS should seek to validate it with other indices of symptom validity and in additional patient population.

The purpose of the present study was to explore the validity of RBS and HHI in predicting performance on the WMT in a VA sample, an instrument found to have greater sensitivity than the TOMM at the recommended cutoffs (Greve, Ord, Curtis, Bianchini, & Brennan, 2008). The present study also sought to delineate whether WMT performance or MMPI-2 validity indices would differentiate individuals involved in a compensation claim.

**Methods**

**Participants**

Data were examined from 194 veterans routinely referred for neuropsychological evaluations who were administered at least the Immediate Recognition (IR) and Delayed Recognition (DR) of the WMT and the MMPI-2. This included those from general clinical referrals, as well as those seen for C&P exams. Twenty participants were excluded due to MMPI-2 invalidity secondary to random responding or true or false response bias (VRIN or TRIN greater than or equal to a T score of 80). MMPI-2 Cannot Say (CNS) was also examined but no cases met the CNS > 30 criterion for removal. Following these exclusions, 174 cases remained. The sample was largely male (157 men, 17 women) with a mean age of 48 years (SD = 11.6) and mean years of education of 12.8 (SD = 2.2). Ethnically, the sample was 68% Caucasian, 26% African American, and 6% Asian/Pacific Islander. Regarding era of military service, the sample consisted of 1% Korean era, 48% Vietnam era, 15% post-Vietnam (predating the Persian Gulf War), 14% Persian Gulf War, and 22% OIF/OEF era. Sixty-six percent of the sample was service-connected for medical disabilities and 29% for psychiatric conditions. Referral sources included mental health clinics in psychiatry, psychology, PTSD, and chemical dependency (37%); primary care (30%); neurology/neurosurgery (13%); compensation (10%); polytrauma (3%); and other (7%).

Although the overall prevalence of neurocognitive response bias within the VA population remains unknown, recent studies, such as Whitney and colleagues (2008) and Axelrod and Schutte (2010), suggest that SVT failure in the context of routine neuropsychological exams is not uncommon. As such, in the current study, an effort was made to broaden the context of compensation-seeking beyond that of the neuropsychological evaluation itself. Similar efforts have been employed related to PTSD by Frueh, Smith, and Barker (1996), who classified individuals as “compensation-seeking” or “noncompensation-seeking” based upon an existing or planned claim for new benefits or an appeal to increase existing benefits. Of the 174 cases in this study, only 21 (12%) were obtained in the context of a compensation disability evaluation. A record review found that 37 additional cases (21% of total sample) were involved in a compensation exam at the same facility within a 9-month timeframe (6 months prior, 3 months post) of their neuropsychological evaluation. The time window is somewhat arbitrary and could be more broadly or narrowly defined; nevertheless, it suggests that participants had an ongoing active claim for compensation for some mental or physical disability that was contemporaneous with the evaluation. The combined group is therefore referred to as the “Compensation-Context” (CC) group, recognizing that the group is not composed solely of those specifically referred for compensation disability evaluations. Of these 58 cases, 35% (N = 20) were seeking compensation for only physical conditions, 12% (N = 7) for only psychological conditions, and 53% (N = 30) for both physical and psychological conditions. Among the diagnoses provided from the neuropsychological evaluations, cognitive and neurological diagnoses were common, although the most prevalent diagnoses provided were psychiatric (77%).

**Instruments and Procedure**

This single-site retrospective study was approved by the institutional review board. All participants were administered the computerized version of the WMT and a complete MMPI-2. Established cutoffs for IR, DR, and Consistency (≤82.5%) were used for the WMT (Green, 2003) and cutoffs of 17 for the RBS (Gervais et al., 2007) and 8 for the HHI (Henry et al., 2006) were used to establish pass/fail status. Gervais and colleagues (2007) found that using a cutoff score of 17 on the RBS corresponded with a specificity of 0.95–1.00 and positive predictive power of 0.77–1.00 in their initial validation sample. Henry and colleagues (2006) found that a cutoff score of 8 accurately classified 85.6% of participants identified as “probable malingering” (p. 788). In the current study, use of the recommended WMT cutoff yielded 89 participants (51%) who passed and 85 participants (49%) who failed the WMT.
Data Analyses

Transformations provided by Gervais, Ben-Porath, Wygant, and Green (2008) allowed for T scores to be generated from raw RBS scores. Scores for MMPI-2 F-family scales and FBS were also analyzed using linear T scores. The HHI was analyzed using raw score totals consistent with scoring criteria provide by G. K. H. Henry (personal communication, May 19, 2009). Means and standard deviations for RBS, HHI, and traditional validity scales for WMT pass/fail groups were generated and Cohen’s d (Cohen, 1988) and confidence intervals were calculated. Point-biserial correlations for mixed dichotomous and continuous variables were conducted to examine the strength of relationships between the criterion WMT groups and MMPI-2 validity and clinical scales.

Similar to the methodology employed by Arbisi and Ben-Porath (1995) and Gervais and colleagues (2007), incremental validity of RBS and HHI was examined through the use hierarchical regression analyses. The present study differed from Arbisi and Ben-Porath (1995) and Gervais and colleagues (2007) in that logistic regression was employed rather than linear regression given the dichotomous DV, noting that significance was determined by χ²-change. Established validity scales were entered in Block 1, and either RBS or HHI was entered in Block 2. Subsequently, the order of entry was reversed to determine whether established validity scales significantly added to the model after RBS or HHI had been entered in Block 1. Using the established cutoffs of 17 for the RBS and 8 for the HHI, we examined the operational characteristics of RBS and HHI in predicting WMT failure and CC group membership, respectively.

Results

As shown in Table 1, the relationships found between WMT pass/fail groups, CC groups, and MMPI-2 validity indicators were modest. Significant correlations with WMT performance were found with the F (Butcher, Ackerman, & Kaleta, 2001), RBS, Fp (Arbisi & Ben-Porath, 1995), and FBS scales. Back Infrequency (Fb; Butcher et al., 2001) and HHI were not found to significantly correlate with WMT performance. WMT performance was not associated with differences in age, education, or ethnicity.

CC group membership was significantly associated with the HHI, FBS, and Fb scales. For this group, F and RBS failed to significantly correlate with membership in the CC group. Regarding demographic variables, membership in the CC group was associated with younger age (r = −.31, p < .001) but not significantly related to education or ethnicity. The relationship between age and CC group membership was not surprising given that a large proportion of younger veterans from the conflicts in Iraq and Afghanistan were scheduled for disability compensation exams during the time the data were collected. Interestingly, WMT failure status was not significantly associated with either involvement in a compensation evaluation (r = .06, p = .42) or CC group status (r = .07, p = .39). Correlations between RBS and other MMPI-2 validity scales were generally quite strong, ranging from r = .73 for FBS to r = .80 for HHI, with only Fp demonstrating a moderate correlation (r = .43). HHI displayed a strongest correlation with FBS (r = .86), which is not surprising given that it is partially derived from FBS and was also highly correlated with RBS (r = .80), MMPI-2 validity scale intercorrelations are found in Table 2.

As shown in Table 3, between-group differences were observed for each MMPI-2 validity scale significantly related to WMT performance: F (t = −2.67, p = .008), RBS (t = −2.61, p = .01), Fp (t = −2.49, p = .01), and FBS (t = −2.06, p = .04). Moderate effect sizes were found, ranging from 0.31 for FBS to 0.40 for F and RBS. Additionally, F and RBS displayed equivalent effect sizes (Cohen’s d = 0.40).

A series of hierarchical logistic regressions were employed to explore whether RBS displayed incremental validity in the prediction of WMT failure above that accounted for by other significantly related scales (i.e., F, Fp, and FBS). Analyses entered F, Fp, or FBS in Block 1 and RBS into Block 2. Additionally, a second series of regressions reversed the order of entry to explore whether established validity scales had incremental validity above that of RBS.

Although each model in Block 2 significantly predicted WMT failure, results suggest that RBS does not offer an additional predictive value beyond that of F (Δχ² = 0.83, p = ns), Fp (Δχ² = 2.88, p = ns), and FBS (Δχ² = 2.55, p = ns). It is also

Table 1. Correlations among WMT and selected MMPI-2 Validity Scales

<table>
<thead>
<tr>
<th></th>
<th>WMT</th>
<th>F</th>
<th>Fb</th>
<th>Fp</th>
<th>FBS</th>
<th>RBS</th>
<th>HHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>.14</td>
<td>.16*</td>
<td>−.01</td>
<td>.16*</td>
<td>.15</td>
<td>.15</td>
<td>.15</td>
</tr>
</tbody>
</table>

Notes: F = Infrequency; Fb = Back Infrequency; Fp = Infrequency-Psychopathology; FBS = Fake Bad Scale; RBS = Response Bias Scale; HHI = Henry–Heilbronner Index; WMT = Word Memory Test; CC = Compensation-Context; MMPI-2 = Minnesota Multiphasic Personality Inventory-2. N = 174.
* p < .05 (two-tailed).
** p < .01 (two-tailed)
worth noting that when entry was reversed, no established validity scale displayed incremental validity over RBS. Results of these analyses are found in Table 4.

Operational characteristics of RBS were examined to assess whether classification accuracy was similar between the validation (Gervais et al., 2007) and the current veteran samples. At the recommended cutoff, RBS had slightly lower sensitivity (0.32) than the validation sample but markedly lower and arguably inadequate specificity (0.81). Positive predictive power (PPP) was 0.61 and negative predictive power (NPP) was 0.55. Examination of alternative cutoff scores indicated that a cutoff score of 19 would provide adequate specificity (0.91) with the cost of further lessening sensitivity (0.24). PPP was 0.71 and NPP was 0.55 using a cutoff score of 19 on RBS.

As shown in Table 5, between-group differences were also evident for each MMPI-2 validity scale significantly related to CC group membership: HHI (t = −3.21, p = .002), FBS (t = −2.45, p = .02), and Fp (t = −2.19, p = .03). Effect sizes were again found to be moderate, ranging from 0.35 for Fp to 0.54 for HHI.

### Table 2. Intercorrelations among RBS, HHI, and established MMPI-2 validity scales

<table>
<thead>
<tr>
<th></th>
<th>RBS</th>
<th>F</th>
<th>Fb</th>
<th>Fp</th>
<th>FBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>.75*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fb</td>
<td>.78*</td>
<td>.85*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fp</td>
<td>.43*</td>
<td>.68*</td>
<td>.55*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FBS</td>
<td>.73*</td>
<td>.50*</td>
<td>.53*</td>
<td>.19</td>
<td></td>
</tr>
<tr>
<td>HHI</td>
<td>.80*</td>
<td>.66*</td>
<td>.67*</td>
<td>.31*</td>
<td>.86*</td>
</tr>
</tbody>
</table>

Notes: F = Infrequency; Fb = Back Infrequency; Fp = Infrequency-Psychopathology; FBS = Fake Bad Scale; RBS = Response Bias Scale; HHI = Henry–Heilbronner Index; MMPI-2 = Minnesota Multiphasic Personality Inventory-2. N = 174.

*p < .01 (two-tailed).

### Table 3. Means, standard deviations, t-tests, effect sizes, and confidence intervals for MMPI-2 validity scales by WMT pass/fail group membership

<table>
<thead>
<tr>
<th></th>
<th>Pass WMT</th>
<th>Fail WMT</th>
<th>t-tests</th>
<th>Cohen's d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>t-value</td>
</tr>
<tr>
<td>F</td>
<td>71.3</td>
<td>19.8</td>
<td>80.0</td>
<td>23.1</td>
</tr>
<tr>
<td>Fb</td>
<td>74.8</td>
<td>23.8</td>
<td>82.2</td>
<td>27.3</td>
</tr>
<tr>
<td>Fp</td>
<td>56.0</td>
<td>14.9</td>
<td>62.3</td>
<td>18.3</td>
</tr>
<tr>
<td>FBS</td>
<td>76.5</td>
<td>16.9</td>
<td>81.6</td>
<td>15.6</td>
</tr>
<tr>
<td>RBS</td>
<td>82.2</td>
<td>17.1</td>
<td>89.4</td>
<td>18.8</td>
</tr>
<tr>
<td>HHI</td>
<td>8.9</td>
<td>4.1</td>
<td>10.0</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Notes: F = Infrequency; Fb = Back Infrequency; Fp = Infrequency-Psychopathology; FBS = Fake Bad Scale; RBS = Response Bias Scale; HHI = Henry–Heilbronner Index; WMT = Word Memory Test; MMPI-2 = Minnesota Multiphasic Personality Inventory-2. All MMPI-2 validity scales shown as linear T scores except HHI shown as a raw score.

### Table 4. Hierarchical logistic regression for prediction of WMT failure by RBS

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>Block</th>
<th>Scale</th>
<th>Model χ² (df)</th>
<th>p-value</th>
<th>χ²-change (df)</th>
<th>p-value</th>
<th>R²</th>
<th>R²-change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>F</td>
<td>7.01 (1)</td>
<td>.01</td>
<td>0.83 (1)</td>
<td>ns</td>
<td>.039</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>RBS</td>
<td>7.84 (2)</td>
<td>.02</td>
<td>.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>RBS</td>
<td>6.70 (1)</td>
<td>.01</td>
<td>1.14 (1)</td>
<td>ns</td>
<td>.038</td>
<td>.006</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>RBS</td>
<td>6.70 (1)</td>
<td>.01</td>
<td>2.41 (1)</td>
<td>ns</td>
<td>.038</td>
<td>.013</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>F</td>
<td>6.23 (1)</td>
<td>.01</td>
<td>2.88 (1)</td>
<td>ns</td>
<td>.035</td>
<td>.016</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>FBS</td>
<td>4.22 (1)</td>
<td>.04</td>
<td>2.55 (1)</td>
<td>ns</td>
<td>.024</td>
<td>.014</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>RBS</td>
<td>6.77 (2)</td>
<td>.03</td>
<td>.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>FBS</td>
<td>6.77 (2)</td>
<td>.03</td>
<td>.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: F = Infrequency; Fp = Infrequency-Psychopathology; FBS = Fake Bad Scale; RBS = Response Bias Scale; WMT = Word Memory Test. Cox and Snell R² utilized.
Another series of hierarchical logistic regressions, which are shown in Table 6, were employed to examine the incremental validity of HHI in the prediction of CC group membership. Either Fb or FBS were entered into Block 1 followed by HHI in Block 2, with the opposite order of entry conducted in a separate series of analyses. Interestingly, HHI displayed significantly improved predictive accuracy of the model when paired with either Fb ($\chi^2 = 5.66, p < .05$) or FBS ($\chi^2 = 4.74, p < .05$). In contrast, when HHI was entered into Block 1 neither Fb nor FBS significantly improved the predictive accuracy of the model when entered in Block 2.

Operational characteristics of HHI in predicting CC group membership were examined given that “substantial external incentive” is Criterion A in the Slick and colleagues (1999) Malingered Neurocognitive Dysfunction diagnostic criteria. At the recommended cutoff of 8, HHI displayed good sensitivity (0.85) but unacceptably low specificity (0.37). Similarly, PPP was quite low (0.40), whereas NPP was quite strong (0.83) at the recommended cutoff. Exploratory analyses examining alternative cutoffs for HHI suggested that the cutoff would need to be raised to an unrealistically high threshold, HHI $\geq 14$, in order to obtain adequate specificity (0.85), and this would also result in a great loss of sensitivity (0.19). Operational characteristics of HHI in predicting WMT performance was not warranted given their non-significant relationship.

**Discussion**

The MMPI-2 is often included in neuropsychological evaluations due to its ability to measure psychopathology and emotional functioning as well as provide measures of response validity. Additional scales have been developed to extend the ability of the MMPI-2 to address response styles including Fp (Arbisi & Ben-Porath, 1995) and FBS (Lees-Haley et al., 1991). The additions of RBS (Gervais et al., 2007) and HHI (Henry et al., 2006) represent extensions of MMPI-2 validity indices as they were derived to detect response bias regarding neurocognitive symptom validity and the exaggeration of somatic and cognitive symptoms, respectively.

Previous studies (Larrabee, 2003; Slick et al., 1996) have found MMPI-2 validity indicators to be positively correlated with the measures of neurocognitive effort. In the current study, scores for RBS, FBS, and F-family scales except for Fb were significantly higher in the WMT fail group. Interestingly, the effect size for RBS generated from the current sample (0.40) was much less robust than the 0.92 reported by Gervais and colleagues (2007). Consistent with the concerns of Nelson and colleagues (2007) regarding RBS’s incremental validity, logistic regression indicated that it did not add incrementally to F,
Fp, or FBS’s modest discrimination of WMT group membership. It is curious that RBS, whose derivation employed the initial 370 items of the MMPI-2, did not show a stronger association with WMT group membership. Regarding operational characteristics, the developers of RBS appropriately selected a conservative cutoff designed to maximize specificity. Unfortunately, the RBS cutoff score of 17 yielded low specificity (0.81) relative to that observed in the initial validation sample. To approach the specificity of 0.95 deemed acceptable in the original validation study, a cutoff of 20 yielding a specificity of 0.93 would have to be applied to our sample. Whitney and colleagues (2008) showed better operational characteristics when using the TOMM as the criterion in their veteran sample, raising the possibility that the TOMM and WMT may differ in their application among veteran patients.

The lack of association of HHI with SVT failure contrasts with the findings of the development and validations studies (Henry et al., 2006, 2008) and Whitney and colleagues (2008), the latter of which was also conducted within a veteran sample. In the present study and Whitney and colleagues (2008), mean HHI scores exceeded the recommended cutoff for individuals passing the administered SVT. This suggests that even VA patients not failing SVTs or identified as compensation-seeking may score high on HHI, potentially resulting in false-positive errors within the VA population. Despite this concern, it is interesting that HHI but not SVT failure was associated with CC in this study, suggesting that HHI may have utility in the prediction of broader aspects of noncredible symptom report than those associated with performance-based SVTs. Although HHI significantly predicted whether the assessment was within a broader CC, its practical use in classifying compensation-seeking was limited. Specifically, HHI had strong sensitivity (0.85) but low specificity (0.37) at the recommended cutoff. Given the importance of avoiding false positives, present results suggest that the recommended cutoff may be too low for some settings and that the limited number of items on HHI provides less opportunity to explore alternative cutoffs.

The less robust findings in this sample raise questions about the ability to generalize RBS and HHI outside of the forensic setting. More specifically, several differences between the current sample and the development and validation samples of RBS and HHI warrant examination and discussion. For example, 87% of the initial validation sample of RBS was involved in a disability claim or litigation, whereas our CC group represented only 33% of the current sample. Similarly, all individuals in the “probable malingering” group of the HHI development sample had failed SVTs and were involved in litigation, whereas the present investigation found that WMT performance and involvement in a compensation claim were not significantly related. Of course, membership in the CC group does not imply that the veteran is seeking to defraud the DVA system, just as litigation status does not indicate the presence of malingering. Rather, it reflects that the veteran has submitted a disability claim and has at least one contemporaneous examination associated with that claim, external incentives that are similar to those expressed in Criterion A for probable malingered neurocognitive dysfunction (Slick et al., 1999). Current results also found that individuals who submitted a disability claim displayed a markedly atypical response pattern on three (i.e., Fb, FBS, HHI) of the eight MMPI-2 validity scales examined. Given that membership in the CC group was broadly defined, there is a wide range of possibilities for the aforementioned findings and suggests that the CC group warrants more study. Although the present sample is largely composed of people involved in claims other than PTSD, a related line of investigations have been conducted with veterans involved in PTSD evaluations and have yielded interesting findings. For example, Frueh, Hamner, Cahill, Gold, and Hamlin (2000) report that veterans with PTSD generally have higher scale elevations on MMPI validity and clinical scales regardless of CC; however, scale elevations were even higher in those classified as “compensation-seeking.” The authors opine that pursuit of disability benefits does not equate to malingering but appears to influence symptom report “either consciously or unconsciously” (p. 865). Related, PTSD symptoms were found to increase between the time that the claim was initiated and the examination was completed (Spoont et al., 2008).

It is also important to broadly consider the context in which veterans generate claims and how being service-connected differs from compensation claims in civil litigation. First, many veterans are encouraged to submit claims by veteran’s advocacy groups, healthcare providers, and peers. This inadvertently creates “noise” in benefits evaluation when considering what precipitated the claim being filed and how this encouragement may influence symptom presentation. Financial incentives are obviously present, but these factors are complex, because the level of compensation and value of accompanying benefits has a curvilinear relationship with the individual’s percentage of service-connection. As such, there is not only motivation to become service-connected but motivation to be service-connected at 100%, because some benefits, such as education benefits for dependents, are available only to individuals who are 100% connected (Frueh et al., 2000). Related, Bianchini, Curtis, and Greve (2006) found evidence of a dose–response relationship between potential compensation and SVT performance in sample of civilian litigant claiming TBI and, as previously described, Spoont and colleagues (2008) found PTSD symptom severity may increase between claim initiation and evaluation. It is also important to consider that less obvious benefits, such as admittance into the VA healthcare system and access to specific forms of medical care, are impacted by a veteran’s service-connection status (Arbisi, Murdoch, Fortier, & McNulty, 2004) and may influence performance.

These aforementioned findings and the non-significant relationship between WMT performance and involvement in a recent compensation evaluation highlight the complexity of assessment within the “hybrid” DVA system, where veterans frequently
utilize VA medical centers for primary medical and mental health services but these frontline services are integrally linked to present, future, and past disability claims. PTSD research examining the relationship between service connection and utilization of mental health services has been mixed (Laffaye, Rosen, Schurr, & Friedman, 2007); however, extensive concern has been expressed that PTSD compensation claims motivate veterans to exaggerate or feign symptoms (Frueh et al., 2000; Frueh, Grubaugh, Elhai, & Buckley, 2007). Although the relationship between disability services and cognitive complaints in a veteran sample has received sparse attention, it is suggested that the interrelationship of services and benefits in the system presents unique challenges for clinicians and measures of response bias. It is also noteworthy that clinicians may be asked to perform C&P evaluations with veterans to whom they have previously provided or in some cases are currently providing clinical services. Whether this impacts symptom report or treatment outcomes is unclear; however, it does present a challenging ethical dilemma related to dual relationships for psychologists and neuropsychologists (American Psychological Association, 2002).

Current findings are tempered by several limitations. First, the present study was archival and patients were recruited based upon the administered measures rather than any specific category of cognitive, medical, or psychiatric complaint. Second, the current study clearly recruited from a different population than individuals recruited for the development and validation samples of RBS and HHI. As such, the present findings do not refute the prior findings of RBS and HHI developers; rather, they suggest that the utility of RBS and HHI may vary by population and/or setting. Third, the present study employed only one SVT as a criterion. Although this is not unlike Whitney and colleagues (2008), who only utilized the TOMM, future studies would benefit from employing multiple free-standing SVTs to explore whether RBS and HHI have variable relationships with different measures of noncredible performance. Future research examining the relationship between utilization of VA services, service connection, and cognitive complaints is greatly needed and would extend conclusions that can be made regarding findings in the present study and in Whitney and colleagues (2008).

The effort to develop an empirically derived, coaching-resistant measure of neurocognitive response bias within the widely used MMPI-2 is a worthy enterprise. The developers of RBS and HHI utilized psychometrically sound approaches in the development of these scales but our findings raise concerns regarding the utility of these scales in all compensation-seeking contexts. As discussed above, sample-specific factors likely played a role in the discrepancy in performance of our sample when compared with the development and validation samples. Further examination of each scale with diverse populations is needed to examine their utility above that demonstrated here.

In SVT research employing known groups, often divergent groups are selected, such as mixed clinical samples for non-compensation seeking groups and litigating samples for compensation seeking groups. In the current study, CC was operationalized within the same institution based on whether or not patients engaged in specific compensation-seeking behaviors within that institution. Broadly speaking, more SVT studies are needed that minimize the differences across compensation-seeking and non-compensation-seeking groups, thereby allowing for closer examination of the influence of secondary gain and whether its value or importance has a differential effect on symptom endorsement and test performance. Such research is especially needed on self-report measures of noncredible symptoms (e.g., MMPI-2, PAI), as such measures often fail to demonstrate the striking differences seen in performance-based SVTs across severe brain injury and compensation seeking groups.

Acknowledgements

This material is the result of work supported with resources and the use of facilities at the Memphis VA Medical Center.

Conflict of Interest

None declared.

References


