The Use of All Three Test of Memory Malingering Trials in Establishing the Level of Effort

Nick M. Wisdom¹,², Whitney L. Brown¹,², David K. Chen³,⁴, Robert L. Collins³,⁴,*

¹Mental Health Care Line, Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX, USA
²Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, TX, USA
³Neurology Care Line, Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX, USA
⁴Department of Neurology, Baylor College of Medicine, Houston, TX, USA

*Corresponding author at: Neurology Care Line, Michael E. DeBakey Veterans Affairs Medical Center, 2002 Holcombe Blvd, Houston, TX 77030, USA. Tel.: +1-713-791-1414; fax: +1-713-794-7631.
E-mail address: robert.collins3@va.gov (R.L. Collins).

Accepted 12 November 2011

Abstract

Assessment of the effort level is an essential step in establishing the internal validity of any neuropsychological evaluation. The use of response bias measures as part of a core battery, however, is less common outside of forensic evaluations. The amount of time needed to administer many of these tests is often cited as a likely explanation for their exclusion from routine neuropsychological evaluations. This study examined all three trials of the Test of Memory Malingering (TOMM) in a large sample (n = 213) of inpatients on an epilepsy monitoring unit with the goal of establishing cut scores for early termination. TOMM Trial 1 demonstrated impressive diagnostic accuracy for determining both adequate and suboptimal levels of effort; various cut scores and classification statistics are presented. The optional Retention trial from the TOMM also increased the hit rate 16% in the detection of poor effort. Clinical implications, limitations, and directions for further research are discussed.

Keywords: Malingering/symptom validity testing; Forensic neuropsychology; Assessment; Professional issues; Norms/normative studies

Introduction

The internal validity of any neuropsychological evaluation can be threatened by the presence of insufficient effort. The potential for monetary compensation or other external incentives can negatively influence the test results because of poor effort even more than the severity of the injury (Binder & Rohling, 1996; Green, Rohling, Lees-Haley, & Allen, 2001). Although the prevalence of suboptimal effort varies by situational factors (e.g., clinical setting, referral question, source of referrals, etc.), there is growing consensus that there is a 40% base rate of performance invalidity in cases where there is the potential for secondary gain (Larrabee, 2003; Mittenberg, Patton, Canyock, & Condit, 2002). This has prompted the American Academy of Clinical Neuropsychology (AACN) and the National Academy of Neuropsychology (NAN) to issue position papers detailing the necessity of including response bias measures in all neuropsychological evaluations (Bush et al., 2005; Heilbronner et al., 2009). Both papers strongly support the use of multiple stand-alone and embedded measures of effort during testing.

There are no response bias tests found among the top 40 most commonly given assessment instruments as reported by practicing neuropsychologists affiliated with Division 40 of the American Psychological Association (APA), the NAN, or the International Neuropsychological Society (INS; Rabin, Barr, & Burton, 2005). Within the specific cognitive domain of memory, the Test of Memory Malingering (TOMM) and Rey 15-item, two response bias measures, come in 19th and 40th, respectively (Rabin et al., 2005). Comparable findings were also observed in a survey conducted several years earlier, suggesting little change despite an increase in the number of publications pertaining to symptom validity tests and response bias over the past...
two decades (Camara, Nathan, & Puente, 2000; Sweet, King, Malina, Bergman, & Simmons, 2002). Recently, a study found that half of NAN members who responded to the survey often or always include measures of effort (Sharland & Gfeller, 2007).

One plausible explanation for the lack of routine response bias administration outside of forensic settings is the amount of time necessary to fully administer many of these measures in general practice settings where time is often limited and malingering occurs at a lower base rate (Gierok, Dickson, & Cole, 2005; Mittenberg et al., 2002; Slick, Tan, Strauss, & Hultsch, 2004). As a result, several recent studies have published normative information for abbreviated versions of a few of the most popular response bias tests (Bauer, O’Bryant, Lynch, McCaffrey, & Fisher, 2007; Doane, Greve, & Bianchini, 2005; Gavett, O’Bryant, Fisher, & McCaffrey, 2005; Greve & Bianchini, 2006; Horner, Bedwell, & Duong, 2006; O’Bryant, Engel, Kleiner, Vasterling, & Black, 2007).

The majority of the aforementioned studies have focused on the TOMM as it is the most frequently given and widely researched measure of effort in the literature (Sharland & Gfeller, 2007; Tombaugh, 1996). The TOMM has demonstrated diagnostic accuracy for detecting insufficient effort and is largely unaffected by age, education, affective distress, or cognitive impairment secondary to most forms of neuropathology (for a review, see O’Bryant et al., 2007). Given the measure’s ubiquity in the field, and in an effort to preserve test security, time will not be spent describing the measure other than to say it is divided into two immediate learning trials and an optional Retention trial given after a 15-min delay (Tombaugh, 1996). Performance below the cutoffs outlined in the manual on Trial 2 and/or the Retention trial is considered a failure as it indicates suboptimal effort.

The TOMM manual does not recommend the use of any cutoffs for Trial 1 to terminate the test early if adequate or inadequate effort is detected. A study conducted by Gavett and colleagues (2005) found that 100% of their referrals for Mild Traumatic Brain Injury involved in active litigation that passed Trial 1 also passed Trial 2 and the Retention trial. This study also used the normative information presented in the TOMM manual to again demonstrate that everyone passing Trial 1 continues to pass the remaining TOMM trials. The findings from this study have been replicated in a general non-litigating clinical sample and in healthy community-dwelling controls, suggesting that the TOMM can confidently be discontinued early if Trial 1 is passed (O’Bryant et al., 2008). In addition, the use of cut scores in identifying suboptimal effort on Trial 1 has also been studied in general clinical samples with consistent findings of impressive diagnostic accuracy (Bauer et al., 2007; Horner et al., 2006; O’Bryant et al., 2007). Finally, the TOMM manual states that the Retention trial is optional, but may help corroborate the results if administered (Tombaugh, 1996, p. 1). A review of the literature yielded only one study examining the increased sensitivity of the TOMM when the Retention trial is administered. Greve and Bianchini (2006) demonstrated a 19% increase in the TOMM’s hit rate in the detection of suboptimal effort when the Retention trial was given. However, their conclusion that the Retention trial should always be administered was challenged on the grounds that 3.1% of the non-litigating cognitively impaired individuals included in the TOMM’s clinically based normative sample failed the Retention trial after passing Trial 2 (Booksh, Aubert, & Andrews, 2007; Tombaugh, 1996). In other words, Booksh and colleagues (2007) argue that the administration of the Retention trial increases the rate of false positives rather than enhancing the TOMM’s sensitivity to detect suboptimal effort.

The purpose of the current study was to replicate the findings from these studies in a unique clinical sample within a Veterans Affairs setting. More specifically, it was hypothesized that nearly all of the individuals passing TOMM Trial 1 would continue to pass the remaining trials. It was also hypothesized that the administration of the optional Retention trial would help identify suboptimal effort in some patients that had previously exerted adequate effort on Trial 2. Finally, this study presents the diagnostic accuracy of TOMM Trial 1 in the detection of suboptimal effort using various cut scores.

Methods

Participants

Data were collected from 213 inpatients (184 men; 29 women), all referred by the Neurology Department for a neuropsychological screening at a large Veterans Affairs Hospital in the southern USA. All the patients were undergoing week-long observation on an epilepsy monitoring unit to establish the presence of epileptic seizures or psychogenic non-epileptic events. The patients ranged in age from 22 to 84 years ($M = 50.3; SD = 13.2$). The sample comprised Caucasian (147; 69%), African American (46; 22%), Hispanic (15; 7%), American Indian (2; 1%), Asian American (1; <1%), and unknown (2; 1%) patients.

Procedure

Both learning trials and the Retention trial of the TOMM were administered to all the patients within the context of a larger neuropsychological screening. Any scores below the standard cut scores found in the TOMM manual on either Trial 2 or the
Retention trial were considered positive (i.e., invalid performance). Consistent with the methods employed in previous studies, the patients were next classified as having put forth adequate or suboptimal effort based on their performance on TOMM Trial 2 and/or the Retention trial (Bauer et al., 2007; O’Bryant et al., 2007). Once the sample was divided, the diagnostic classification statistics for TOMM Trial 1 were calculated. More specifically, the sensitivity (i.e., ability to detect true positives) and specificity (i.e., accurately identifying true negatives) values were calculated for a range of cut scores. The positive predictive value (the proportion of patients positively identified by TOMM Trial 1 that were actually putting forth inadequate effort) and the negative predictive value (the proportion of patients with negative findings on TOMM Trial 1 that were actually giving adequate effort) were calculated while taking into account various base rates of suboptimal effort that have been observed in different clinical settings and populations (Rosenfeld, Sands, & Van Gorp, 2000).

**Results**

No significant differences were noted between age, gender, or ethnic groups on any of the TOMM trials administered. Out of all 213 patients, a total of 22 (10.3%) failed TOMM Trial 2 and/or the Retention trial. The remaining patients were considered as having put forth adequate effort on the TOMM.

**TOMM Trial 1**

A total of 142 patients (66.7%) passed Trial 1 using the same cut score as the one recommended in the TOMM manual for Trial 2 and the Retention trial. Of those 142 patients, 141 (99.3%) went on to pass Trial 2 and the Retention trial. The only patient that did not continue to pass the remaining trials was 1 point below the cut score on Trial 2 and within normal limits on the Retention trial.

**Retention Trial**

A total of 194 patients (91%) passed Trial 2 of the TOMM. Of those 194 patients, 191 (98.5%) also passed the Retention trial. Two of the patients who failed were 1 point below the cut score and the remaining patient was 6 points below the cut score on the Retention trial. As a result of these three additional patients failing the Retention trial, the TOMM’s theoretical hit rate of detecting suboptimal effort in this sample increased 16%.

**Discriminant Validity of TOMM Trial 1**

Table 1 presents the diagnostic classification statistics for the TOMM Trial 1 in the detection of suboptimal effort. While reviewing the table, it is important to remember that these cut scores were derived from patients that gave adequate (n = 191) and suboptimal (n = 22) effort on Trial 2 and/or the Retention trial. In other words, these cut scores represent the number that needs to be exceeded in order to be considered as putting forth good effort. Using a cut score of \( \leq 39 \) on TOMM Trial 1 to establish suboptimal effort yields the highest ratio of sensitivity and specificity. Various base rates are also presented in this table so that the other cut scores can be used depending on the clinical setting and/or the population being evaluated.

**Discussion**

The use of response bias measures is an essential component of any neuropsychological evaluation in that they help establish the level of effort exerted by the examinee. While we encourage adherence to standardized instructions and test administration, it is important to recognize the impracticality of dedicating a substantial amount of a test battery to measures of response bias when there is only modest reason to suspect poor effort given the context of the referral (e.g., general outpatient vs. medicolegal). As a result, a number of studies have started documenting early discontinue criteria for some of the most popular response bias measures, particularly the TOMM.

Similar to previous findings in the literature (Gavett et al., 2005; O’Bryant et al., 2008), this study found that 99.3% of the patients scoring above the cut score on Trial 1 continued to pass the remaining trials of the TOMM. It is reasonable to conclude that early termination of the TOMM after establishing good effort on Trial 1 is extremely unlikely to result in a missed patient that would have put forth suboptimal effort on either of the remaining trials. Also, consistent with the study conducted by Greve and Bianchini (2006), administration of the Retention trial yielded three additional patients considered as having put forth invalid performances per the recommendations in the TOMM manual. This represents a 16% and 19% increase in the TOMM’s theoretical hit rate in the detection of poor effort in the current study and Greve and Bianchini’s study (2006),
respectively. The concern that this percentage increase is the result of false positives as opposed to suboptimal effort on the Retention trial is still relevant (Booksh et al., 2007), but at least one patient in this study performed significantly below the cut score after putting forth good effort on the first two trials. Finally, this study presents diagnostic classification statistics for Trial 1 from the TOMM and findings are remarkably congruent with previous studies. This study found that using a cut score of \( \leq 39 \) was very likely to result in the failure of TOMM Trial 2 and/or the Retention trial. Bauer and colleagues (2007) and O’Bryant and colleagues (2007) each maximized the ratio between sensitivity and specificity using TOMM Trial 1 cut scores of \( \leq 37 \) and \( \leq 38 \), respectively. Horner and colleagues (2006) recommended using a cut score of \( \leq 35 \), but the authors acknowledge that this is a very conservative approach in light of the ramifications of false-positive errors. Regardless of these recommendations, individual clinicians should select the most appropriate cut score for their own clinical setting after controlling for the base rate of response bias failure observed within their sample. For example, the positive predictive value of the TOMM decreases as the base rate of suboptimal effort decreases. This is because the likelihood of a positive finding is less common and statistically less likely in samples with a lower base rate. Similarly, the negative predictive value decreases in samples with a high base rate of suboptimal effort (e.g., forensic settings) because it statistically occurs more frequently and is harder to rule out. There are several limitations present in this study that need to be considered when evaluating the results. The primary criticism is that the TOMM was the only measure used to divide the sample into adequate and inadequate effort groups and then the diagnostic classification statistics were calculated for TOMM Trial 1. As a result, the findings from the data presented in Table 1 should only be used to predict performance on TOMM Trial 2 and the Retention trial and do not generalize to other measures of response bias. Another limitation is the relatively low base rate of performance invalidity in this sample (10.1%) as determined by performance on the TOMM. Settings with higher base rates of TOMM failure are statistically more likely to have variability in scores and/or unusual presentations not detectable in this or prior study samples. This is a minor concern as it is unlikely that neuropsychologists working in forensic settings would abbreviate their administration of response bias measures. Regardless, future studies should attempt to replicate these findings using a sample with a higher base rate of TOMM failure. In addition, more studies are needed comparing the concordance rate of failure when using multiple symptom validity tests within the same sample (Greiffenstein, Greve, Bianchini, & Baker, 2008). Finally, there is a need in the literature for an updated survey study examining the frequency of response bias testing as part of a

<table>
<thead>
<tr>
<th>Cut score</th>
<th>SE</th>
<th>SP</th>
<th>Base rate (0.40)</th>
<th>Base rate (0.30)</th>
<th>Base rate (0.20)</th>
<th>Base rate (0.10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>PPV</td>
<td>SPV</td>
<td>PPV</td>
<td>SPV</td>
</tr>
<tr>
<td>30</td>
<td>0.27</td>
<td>1.00</td>
<td>1.00</td>
<td>0.67</td>
<td>1.00</td>
<td>0.76</td>
</tr>
<tr>
<td>31</td>
<td>0.32</td>
<td>1.00</td>
<td>1.00</td>
<td>0.69</td>
<td>1.00</td>
<td>0.77</td>
</tr>
<tr>
<td>32</td>
<td>0.36</td>
<td>0.98</td>
<td>0.93</td>
<td>0.71</td>
<td>0.90</td>
<td>0.79</td>
</tr>
<tr>
<td>33</td>
<td>0.41</td>
<td>0.98</td>
<td>0.94</td>
<td>0.75</td>
<td>0.91</td>
<td>0.82</td>
</tr>
<tr>
<td>34</td>
<td>0.50</td>
<td>0.98</td>
<td>0.95</td>
<td>0.77</td>
<td>0.92</td>
<td>0.84</td>
</tr>
<tr>
<td>35</td>
<td>0.55</td>
<td>0.97</td>
<td>0.92</td>
<td>0.76</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>36</td>
<td>0.55</td>
<td>0.97</td>
<td>0.93</td>
<td>0.78</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>37</td>
<td>0.59</td>
<td>0.96</td>
<td>0.92</td>
<td>0.82</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>38</td>
<td>0.68</td>
<td>0.93</td>
<td>0.88</td>
<td>0.86</td>
<td>0.83</td>
<td>0.90</td>
</tr>
<tr>
<td>39</td>
<td>0.77</td>
<td>0.91</td>
<td>0.85</td>
<td>0.86</td>
<td>0.79</td>
<td>0.90</td>
</tr>
<tr>
<td>40</td>
<td>0.77</td>
<td>0.88</td>
<td>0.81</td>
<td>0.85</td>
<td>0.73</td>
<td>0.90</td>
</tr>
<tr>
<td>41</td>
<td>0.80</td>
<td>0.84</td>
<td>0.79</td>
<td>0.91</td>
<td>0.70</td>
<td>0.94</td>
</tr>
<tr>
<td>42</td>
<td>0.88</td>
<td>0.80</td>
<td>0.75</td>
<td>0.91</td>
<td>0.65</td>
<td>0.94</td>
</tr>
<tr>
<td>43</td>
<td>0.96</td>
<td>0.78</td>
<td>0.74</td>
<td>0.97</td>
<td>0.65</td>
<td>0.98</td>
</tr>
<tr>
<td>44</td>
<td>0.96</td>
<td>0.74</td>
<td>0.71</td>
<td>0.97</td>
<td>0.61</td>
<td>0.98</td>
</tr>
<tr>
<td>45</td>
<td>0.96</td>
<td>0.67</td>
<td>0.66</td>
<td>0.96</td>
<td>0.55</td>
<td>0.98</td>
</tr>
<tr>
<td>46</td>
<td>0.96</td>
<td>0.60</td>
<td>0.62</td>
<td>0.96</td>
<td>0.51</td>
<td>0.97</td>
</tr>
<tr>
<td>47</td>
<td>1.00</td>
<td>0.52</td>
<td>0.58</td>
<td>1.00</td>
<td>0.47</td>
<td>1.00</td>
</tr>
<tr>
<td>48</td>
<td>1.00</td>
<td>0.40</td>
<td>0.53</td>
<td>1.00</td>
<td>0.42</td>
<td>1.00</td>
</tr>
<tr>
<td>50</td>
<td>1.00</td>
<td>0.25</td>
<td>0.47</td>
<td>1.00</td>
<td>0.36</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Notes: SE = sensitivity \[\text{true positives}/(\text{true positives} + \text{false negatives})\]; SP = specificity \[\text{true negatives}/(\text{true negatives} + \text{false positives})\]; PPV = positive predictive value \[\text{SE \times base rate}}{	ext{SE \times base rate} + (1 - \text{SP}) \times (1 - \text{base rate})}\]; NPV = negative predictive value \[\text{SP \times (1 - \text{base rate})}/(\text{SP \times (1 - \text{base rate})} + (1 - \text{SE}) \times \text{base rate})\].
routine neuropsychological evaluation. It is likely that many of the aforementioned studies do not reflect the increasing popularity of newer, well-validated response bias measures.

References


