Comparison of Performance of the Test of Memory Malingering and Word Memory Test in a Criminal Forensic Sample

Rachel L. Fazio1,*, James Forrest Sanders1, Robert L. Denney1,2

1The School of Professional Psychology, Forest Institute, Springfield, MO, USA
2Neuropsychological Associates of Southwest Missouri, Springfield, MO, USA

*Corresponding author at: The School of Professional Psychology, Forest Institute, 2885 W Battlefield Rd, Springfield, MO 65807, USA. Tel.: +1-505-274-0828; fax: +1-417-893-7842. E-mail address: rlfazio@gmail.com (R. L. Fazio).

Accepted 2 April 2015

Abstract

Compared with the amount of neuropsychological literature surrounding response bias in civil litigation, there is little regarding criminal cases. This study adds to the criminal forensic neuropsychological literature by comparing the Test of Memory Malingering (TOMM) and the Word Memory Test (WMT) in a criminal forensic setting utilizing a criterion-groups design. Subjects were classified into two groups based on their performance on at least two other freestanding performance validity tests. The WMT demonstrated good sensitivity (95.1%) but poor specificity (68.4%) when Genuine Memory Impaired Profiles (GMIPs) were not considered. Inclusion of GMIPs reduced the sensitivity to 56.1% but increased the specificity to 94.7%. The TOMM evidenced better sensitivity but poorer specificity than the WMT with GMIPs. Conjoint use of the tests was also considered. Receiver operating characteristics and other classification statistics for each measure are presented. Results support the use of these measures in a criminal forensic population.

Keywords: Forensic Neuropsychology; Malingering; Performance Validity; Response Bias; Test of Memory Malingering; Word Memory Test

Introduction

The opinion of forensic neuropsychologists regarding the relevance of neuropathological sequelae to civil damages and/or criminal competency and responsibility is beginning to comprise a significant portion of expert testimony in courts (Farkas, Rosenfeld, Robbins, & van Gorp, 2006; Heilbronner, 2004). In a forensic context, incentive to malinger neurocognitive dysfunction (MND) is great (Ardolf, Denney, & Houston, 2007; Iverson, 2006; Iverson & Binder, 2000; Mittenberg, Patton, Canyock, & Condit, 2002; Rogers, 2008). Assessment of effort, response bias, and malingering is crucial to ensure valid test results (Bianchini, Mathias, & Greve, 2001; Iverson, 2006). So critical is the issue that professional societies and governing bodies are citing the assessment of effort and malingering as standard of care (Bush et al., 2005; Heilbronner, Sweet, Morgan, Larrabee, & Millis, 2009).

Current state of the art in detecting MND incorporates neuropsychological assessment, self-report data, and behavioral observation (Iverson & Binder, 2000; Nies & Sweet, 1994; Pankratz and Binder, 1997; Slick, Sherman, & Iverson, 1999). The utilization of neuropsychological test data is a particularly powerful and reliable method (Bianchini et al., 2001). Freestanding performance validity tests (PVTs) that rely on forced-choice methodology and normative-based reference scores have become the “gold standard” (Slick et al., 1999). Among the various PVTs, the Word Memory Test (WMT; Green, 2005) and the Test of Memory Malingering (TOMM; Tombaugh, 1996) are two of the most researched, well validated, and widely used measures of symptom validity and response bias (Greve, Ord, Curtis, Bianchini, & Brennan, 2008). Though often studied, little research exists comparing TOMM and WMT performance in “known-group” designs, wherein the clinical application of these measures is best assessed (Gervais, Rohling, Green, & Ford, 2004; Greiffenstein, Baker, & Gola, 1994; Rogers, 2008; Tombaugh, 1996). A known-group design uses a criterion classification system, typically other tests and/or diagnostic system (e.g., Slick et al., 1999) to determine how actual examinees are classified. In this context, the issue is malingering or not malingering. Since the classification accuracy of each case individually is not truly known, we prefer the term criterion-group design over known-group design.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Currently, the only criterion-group comparison of the TOMM and WMT was conducted by Greve and colleagues (2008). The authors employed results of 109 examinees with purported traumatic brain injury. Results suggested the WMT offered much better sensitivity but poorer specificity than the TOMM, signifying that each measure holds different strengths and weaknesses. However, they did not incorporate the full WMT profile analysis in the study, a method recommended by the author to decrease false positives (Green, 2005). The researchers were able to improve on classification accuracy by utilizing the TOMM and the WMT together, which offers support to current guidelines recommending the utilization of several PVTs in clinical practice.

Forensic neuropsychology has been practiced primarily within the civil arena, where research supports the use of the TOMM and the WMT. Neuropsychological assessment is increasingly requested in criminal cases. Currently, no studies exist comparing the performance of these measures in criminal forensic settings. Most studies of response bias and malingering in criminal forensic settings have focused on malingering of psychiatric symptoms. Among extant literature evaluating MND in a criminal forensic setting, however, two primary differences can be noted. First, studies have suggested elevated base rates of MND in criminal forensic settings when compared with civil settings (Ardolf et al., 2007; Heaton, Smith, Lehman, & Vogt, 1978). Second, some authors have noted elevated rates of suppression on forced-choice measures (Frederick & Denney, 1998). Assuming the TOMM and the WMT perform similarly across criminal and civil settings potentially carries costly outcomes, such as an undue loss of liberty or justice for the guilty (Tombaugh, 1996; Weinborn, Orr, Woods, Conover, & Feix, 2003). These noted differences between MND in criminal and civil forensic settings and the potential detriment of assuming equivalent performance across settings supports the rationale for evaluating and comparing the performance characteristics of the TOMM and the WMT in a criminal forensic setting. In addition to informing clinicians working in a criminal forensic setting about TOMM and WMT performance, evaluating and comparing the TOMM and WMT via criterion-group design supports legal admissibility of these instruments into testimony according to the Daubert standard (Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993).

The purpose of this study was to assess and compare the diagnostic accuracies of the TOMM and the WMT in a criminal forensic setting while using the profile analysis method recommended by the WMT author (Green, 2005). Following a priori classification, clinical accuracy was assessed for each measure by way of Receiver Operator Characteristic curve (ROC) analysis. Area under the curve (AUC), sensitivity, specificity, and other relevant clinical statistics were calculated for the TOMM and the WMT individually and conjointly. Additionally, we compared the predictive accuracy of the WMT with and without the full profile analysis.

Methods

Examinees

Examinees were 225 males drawn from an archival database maintained at a maximum security forensic hospital in the Midwest. All examinees underwent neuropsychological evaluation as part of a forensic evaluation or on referral for suspected neurocognitive dysfunction as part of a medical evaluation. All defendants were court-ordered for mental health examination pertaining to their legal competency to proceed, where neurocognitive status was at issue. Most defendants were referred as part of an evaluation for competency to stand trial. Some, however, were evaluated as part of treatment for restoration to competency. The remainder of the sample included sentenced inmates referred for neuropsychological evaluation because of a documented history of neurologic injury or disease or due to concern over neurologic/neuropsychological functioning.

An examinee was included in the analysis if the individual was administered the TOMM and the WMT and at least two other freestanding measures of performance validity. One-hundred and nine of the original 225 examinees satisfied criteria. This resulted in a total sample that was 43.1% Caucasian, 31.2% African American, 12.8% Hispanic, 8.3% Native American, 3.7% Asian, and 0.9% biracial. The average age was 40.1 years (SD = 13.4), average education was 10.5 years (SD = 2.9), and average FSIQ was 76.5 (SD = 16.7). The low mean FSIQ in this sample is likely an underestimate related to performance characteristics of individuals in the study and is further addressed below.

Procedures

Criterion-group classification. Examinees were considered positive for probable malingered neurocognitive dysfunction (+MND) if they met criteria for poor effort on at least two freestanding tests of performance validity, not including the WMT or the TOMM. Two positive indicators have been suggested as sufficient indication of probable MND (Larrabee, Greiffenstein, Greve, & Bianchini, 2007). Individual’s performances were presumed valid (–MND) if they did not meet criteria for poor effort on any of the freestanding tests of performance validity, again excluding the WMT or the TOMM.
Criterion-group results. A total of 41 (51.9%) examinees satisfied the criteria for classification into the +MND group and 38 were grouped as −MND. The remaining individuals had failed one freestanding PVT; these participants were eliminated from further analyses to reduce the rate of potential false positives and false negatives. The +MND group was 26.8% Caucasian, 34.1% African American, 29.3% Hispanic, and 9.8% Native American. Examinees in the +MND group ranged in age from 22 to 67 with a mean age of 39.3 (SD = 11.0). Education was recorded as the number of years completed and averaged 9.7 years (SD = 2.9). The mean IQ for the group was 63.6 (SD = 13.3). Purported diagnoses at the time of evaluation included mild traumatic brain injury (n = 10), moderate traumatic brain injury (n = 6), unspecified traumatic brain injury (n = 4), substance abuse (n = 4), learning disability/mental retardation (n = 4), gunshot wound to the head (n = 2), and a variety of other diagnoses.

The −MND group was 50.0% Caucasian, 31.6% African American, 2.6% Hispanic, 7.9% Native American, and 7.9% Asian. Examinees in the −MND group ranged in age from 20 to 82 with a mean of 40.1 (SD = 14.1). They averaged 11.1 (SD = 3.0) years of completed education. The mean IQ was 87.8 (SD = 12.2). The −MND group included presenting diagnoses of mild traumatic brain injury (n = 7), moderate traumatic brain injury (n = 2), unspecified traumatic brain injury (n = 4), psychiatric diagnoses (n = 3), seizure disorder (n = 1), substance abuse (n = 1), gunshot wound to the head (n = 3), HIV (n = 1), stroke (n = 4), and various other diagnoses.

A priori groups were created based on aforementioned criteria in order to assess and compare TOMM and WMT performance. As per the test manuals, a score of less than 90% on TOMM Trial 2 or the Retention Trial was considered indicative of negative response bias. WMT scores of less than 83% on Immediate Recognition (IR), Delayed Recognition (DR), or Consistency (CNS) trials were considered indicative of negative response bias, pending further analysis. Cases with WMT scores below 83% on IR, DR, or CNS were further analyzed for the presence of genuine memory impairment using the Genuine Memory Impairment Profile (GMIP; Green, 2005).

The GMIP is a more recent addition to the WMT and serves as an additional interpretative strategy for assessing failed WMT performances suggestive of poor effort (i.e., below manual cutoffs, but not statistically significant below random performances on IR, DR, or CNS) which result from genuine significant cognitive impairment rather than poor effort. The GMIP is a three-step algorithm. The algorithm compares performance on WMT effort indices (IR, DR, and CNS) to the test’s more difficult memory indices, which include Multiple Choice (MC), Paired Associates (PA), and Free Recall (FR). If the difference between the mean score of effort indices (IR, DR, and CNS) is 30 or greater than the mean score of memory indices (MC, PA, and FR) then the individual’s performance is further considered for genuine memory impairment in the second step of the algorithm. Step two requires that scores on actual memory measures PA and FR follow a performance pattern clinically congruent with significant cognitive impairment. That is, PA scores are greater than scores on FR. Finally, should individuals pass these initial two steps, scores on IR and DR (effort indices) are compared with the individuals FR score (actual memory score) with the expectation that effort indices should always be higher than more difficult memory measures in the presence of genuine memory impairment. If all three of the aforementioned qualifications exist in the presence of below criterion (but not statistically below random) failing scores on effort indices, the individual’s performance is congruent with that expected of persons with genuinely significant cognitive impairment and not poor effort or malingering. Such an analysis is particularly important in settings where potentially severe neuropathological conditions are present (e.g., severe traumatic brain injury, dementing conditions, etc.). Subjects with WMT results consistent with severe cognitive impairment according to GMIP were not included in the +MND group.

Materials. Freestanding validity tests used to classify examinees were the Computerized Assessment of Response Bias (CARB; Allen, Conder, Green, & Cox, 1997), Validity Indicator Profile (VIP; Frederick, 2003), Rey Fifteen Item Test (Boone, Salazar, Lu, Warner-Chacon, & Razani, 2002; Rey, 1958), Victoria Symptom Validity Test (VSVT; Slick, Hopp, Strauss, & Thompson, 1997), and the Medical Symptom Validity Test (MSVT; Green, 2004). The MSVT GMIP profile was utilized in classification to maximize accuracy of a priori classification. Not all tests were available for all participants; see Table 1 for a depiction of the percentage of participants administered each instrument by group. As seen in Table 1, the CARB and the Rey Fifteen Item Test were the two most frequently administered tests. Of the total sample, 83.5% were administered these two tests together; however, only 8.9% of the sample received these two tests alone, with the majority being administered one or more additional freestanding PVT.

Statistical analysis. Distribution of age, education, and IQ for the groups was non-normal. Due to the non-normal distributions, Mann–Whitney $U$ tests were employed to assess significant differences between +MND and −MND groups. For the same reasons, Mann–Whitney $U$ tests were utilized in assessing significant between group differences on the TOMM and WMT indices of response bias. TOMM and WMT indices were subjected to ROC analysis. Sensitivity and specificity for all TOMM and WMT indices were reported. For assessing clinical accuracy at the individual level, positive and negative predictive powers were calculated at a variety of base rates. Lastly, joint classification accuracies were conducted to assess detection of response bias when the TOMM and the
WMT are used conjointly. Joint classification was assessed for both of the following scenarios: when any TOMM or WMT measure indicated negative response bias and when any TOMM and WMT measure agreed on the presence of response bias.

**Results**

Statistical comparisons revealed no difference in education between groups ($U = 637, p = .16$). The two groups also did not differ significantly on age ($U = 771.5, p = .94$). Conversely, there was a significant difference regarding the distribution of race/ethnicity between the groups ($\chi^2 = 14.6, p = .005$). Inspection of the data revealed that there was a larger proportion of Caucasians in the $-\text{MND}$ group, whereas the $+\text{MND}$ group had more evaluators who were Hispanic. IQ for the $+\text{MND}$ group was significantly lower than IQ for the $-\text{MND}$ group ($U = 155, p < .001$). Given the relationship of individual psychometric measures to effort, negative response bias, and malingered, differences in IQ were expected (Demakis et al., 2000; Green, Rohling, Lees-Haley, & Allen, 2001). Additionally, research demonstrates that IQ tests are quite susceptible to intentional suppression of performance (Johnstone & Cooke, 2003). Groups were also assessed to determine if a difference in the number of tests administered existed between the groups, that is, to see if the $+\text{MND}$ group was primarily positive due to having been administered more tests. The $-\text{MND}$ group was administered an average of 3.4 freestanding PVTs, whereas the $+\text{MND}$ group was administered 3.8. This was not significantly different ($U = 621, p = .104$).

Differences between $+\text{MND}$ and $-\text{MND}$ PVT performances were assessed at the group level. Between groups, differences for TOMM and WMT indices of response bias were conducted using the Mann–Whitney $U$ test due to differences between data distributions in the two groups. Data in the $+\text{MND}$ typically had a flat distribution, while the $-\text{MND}$ group was entirely negatively skewed. Group differences emerged for all TOMM and WMT measures of response bias. Table 2 displays means, standard deviations, and relevant test statistics.

AUCs were identified for individual WMT indices of response bias. As seen in Table 3, the Immediate Recognition trial produced a good AUC ($AUC = 0.875, SE = 0.039, p < .001$). CNS discriminated between the groups slightly better ($AUC = 0.889, SE = 0.033, p < .001$). The discriminative ability of the Delayed Recognition trial appeared to be better than IR or CNS with an AUC of 0.901 ($SE = 0.036, p < .001$).

### Table 1. Percentage of groups administered each performance validity test

<table>
<thead>
<tr>
<th>Index</th>
<th>$-\text{MND}$ % Admin</th>
<th>$+\text{MND}$ % Admin</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSVT</td>
<td>44.7</td>
<td>53.7</td>
</tr>
<tr>
<td>CARB</td>
<td>84.2</td>
<td>90.2</td>
</tr>
<tr>
<td>Rey 15 Item Test</td>
<td>84.2</td>
<td>95.1</td>
</tr>
<tr>
<td>VIP Verbal Subtest</td>
<td>57.9</td>
<td>61.0</td>
</tr>
<tr>
<td>VIP Non-verbal Subtest</td>
<td>63.2</td>
<td>70.7</td>
</tr>
<tr>
<td>VSVT</td>
<td>7.9</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Notes: $\text{MND} = \text{Malingered Neurocognitive Dysfunction}; \text{CARB} = \text{Computerized Assessment of Response Bias}, <85\%$ on any trial (Allen et al., 1997); MSVT = Medical Symptom Validity Test, $<85\%$ on IR, DR, or CNS, Genuine Memory Impaired Profiles excluded (Green, 2004); Rey 15 Item Test, recall $<9$ or $<20$ total score with recognition (Boone et al., 2002; Rey, 1958); VIP = Validity Indicator Profile, irrelevant or suppressed profile (Frederick, 2003); VSVT = Victoria Symptom Validity Test, questionable or invalid on Easy, Hard, or Total scores (Slick et al., 1997).

### Table 2. Between group differences on TOMM and WMT indices

<table>
<thead>
<tr>
<th>Index</th>
<th>$-\text{MND}$ ($M \pm SD$)</th>
<th>$+\text{MND}$ ($M \pm SD$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>90.6 (11.7)</td>
<td>62.9 (21.2)</td>
</tr>
<tr>
<td>DR</td>
<td>91.2 (11.6)</td>
<td>55.7 (23.4)</td>
</tr>
<tr>
<td>CNS</td>
<td>86.5 (13.4)</td>
<td>60.4 (15.1)</td>
</tr>
<tr>
<td>TOMM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 2</td>
<td>48.2 (5.4)</td>
<td>34.3 (15.2)</td>
</tr>
<tr>
<td>Retention</td>
<td>48.2 (5.6)</td>
<td>32.6 (15.4)</td>
</tr>
</tbody>
</table>

Notes: $M = \text{mean}; SD = \text{standard deviation}; \text{MND} = \text{Malingered Neurocognitive Dysfunction}. WMT = \text{Word Memory Test}; IR = \text{Immediate Recall \%}; DR = \text{Delayed Recall \%}; CNS = \text{Consistency \%}; TOMM = \text{Test of Memory Malingering}, raw scores. All $p < .001$. 

---

R. L. Fazio et al. / Archives of Clinical Neuropsychology 30 (2015); 293–301

Downloaded from https://academic.oup.com/acn/article-abstract/30/4/293/5500/Comparison-of-Performance-of-the-Test-of-Memory by guest on 04 October 2017
AUCs were calculated for TOMM Trial 2 and the Retention Trial. Trial 2 evidenced an AUC of 0.793 (\(SE = 0.052, p < .001\)) and the Retention Trial evidenced an AUC of 0.839 (\(SE = 0.046, p < .001\)). Thus, indices of response bias on the TOMM produced good discriminatory performance, but scores were not quite as high as those of the WMT.

Clinically relevant performance characteristics are expressed in sensitivity, specificity, likelihood ratios, and positive and negative predictive power. As evidenced in Table 4, the TOMM had greater sensitivity than the WMT when GMIP was employed, whereas the WMT was highly specific. When using agreement between the TOMM and the WMT, sensitivity was reduced to 48.8% but specificity was 100%. For purposes of comparison with previously published studies, the same performance characteristics are also presented for the WMT without consideration of the GMIP profile and conjoint classification accuracies in Table 4.

Table 5 presents the positive and negative predictive powers for the tests at clinically relevant base rates. At the base rates of MND seen in a criminal forensic neuropsychology setting (Ardolf et al., 2007; i.e., 50% or greater), the WMT held the advantage.

### Table 3. AUC for TOMM and WMT indices

<table>
<thead>
<tr>
<th>Index</th>
<th>AUC</th>
<th>SE</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>0.875</td>
<td>0.039</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DR</td>
<td>0.901</td>
<td>0.036</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>CNS</td>
<td>0.889</td>
<td>0.037</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>TOMM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 2</td>
<td>0.793</td>
<td>0.052</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Retention</td>
<td>0.839</td>
<td>0.046</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Notes: AUC = Area under the curve; SE = standard error; WMT = Word Memory Test; IR = Immediate Recall; DR = Delayed Recall; CNS = Consistency; TOMM = Test of Memory Malingering.

### Table 4. Sensitivity, specificity, false-positive and false-negative error rate, misclassification rate, and negative and positive likelihood ratios

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Sens</th>
<th>Spec</th>
<th>FP</th>
<th>FN</th>
<th>MR</th>
<th>(-LR)</th>
<th>(+LR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT</td>
<td>56.1</td>
<td>94.7</td>
<td>5.3</td>
<td>43.9</td>
<td>25.3</td>
<td>2.2</td>
<td>10.7</td>
</tr>
<tr>
<td>TOMM</td>
<td>68.2</td>
<td>86.8</td>
<td>13.2</td>
<td>31.7</td>
<td>22.8</td>
<td>2.7</td>
<td>5.2</td>
</tr>
<tr>
<td>TOMM or WMT</td>
<td>75.6</td>
<td>81.6</td>
<td>18.4</td>
<td>24.4</td>
<td>21.5</td>
<td>3.3</td>
<td>4.1</td>
</tr>
<tr>
<td>TOMM and WMT</td>
<td>48.8</td>
<td>100.0</td>
<td>0.0</td>
<td>51.2</td>
<td>26.6</td>
<td>2.0</td>
<td>—</td>
</tr>
<tr>
<td>WMT (no GMIP)</td>
<td>95.1</td>
<td>68.4</td>
<td>31.6</td>
<td>4.9</td>
<td>17.7</td>
<td>14.0</td>
<td>3.0</td>
</tr>
<tr>
<td>TOMM or WMT</td>
<td>95.1</td>
<td>63.2</td>
<td>36.8</td>
<td>4.9</td>
<td>20.1</td>
<td>12.9</td>
<td>2.6</td>
</tr>
<tr>
<td>TOMM and WMT</td>
<td>68.3</td>
<td>92.1</td>
<td>7.9</td>
<td>31.7</td>
<td>20.3</td>
<td>2.9</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Notes: WMT = Word Memory Test; TOMM = Test of Memory Malingering; GMIP = Genuine Memory Impaired Profile; Sens = sensitivity; Spec = specificity; FP = false positives; FN = false negatives; MR = misclassification rate; \(-LR\) = negative likelihood ratio; \(+LR\) = positive likelihood ratio.

### Table 5. Positive and negative predictive values of WMT and TOMM indices at various base rates

<table>
<thead>
<tr>
<th>NRB indicators</th>
<th>Base rates</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT PPV</td>
<td></td>
<td>0.542</td>
<td>0.727</td>
<td>0.816</td>
<td>0.820</td>
<td>0.914</td>
<td>0.941</td>
<td>0.961</td>
</tr>
<tr>
<td>WMT NPV</td>
<td></td>
<td>0.951</td>
<td>0.896</td>
<td>0.834</td>
<td>0.763</td>
<td>0.683</td>
<td>0.589</td>
<td>0.480</td>
</tr>
<tr>
<td>TOMM PPV</td>
<td></td>
<td>0.366</td>
<td>0.565</td>
<td>0.690</td>
<td>0.778</td>
<td>0.838</td>
<td>0.886</td>
<td>0.924</td>
</tr>
<tr>
<td>TOMM NPV</td>
<td></td>
<td>0.961</td>
<td>0.916</td>
<td>0.865</td>
<td>0.804</td>
<td>0.733</td>
<td>0.646</td>
<td>0.540</td>
</tr>
<tr>
<td>TOMM or WMT PPV</td>
<td></td>
<td>0.313</td>
<td>0.506</td>
<td>0.638</td>
<td>0.732</td>
<td>0.804</td>
<td>0.860</td>
<td>0.905</td>
</tr>
<tr>
<td>TOMM or WMT NPV</td>
<td></td>
<td>0.968</td>
<td>0.930</td>
<td>0.886</td>
<td>0.834</td>
<td>0.770</td>
<td>0.690</td>
<td>0.589</td>
</tr>
<tr>
<td>TOMM and WMT PPV</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>TOMM and WMT NPV</td>
<td></td>
<td>0.946</td>
<td>0.886</td>
<td>0.820</td>
<td>0.745</td>
<td>0.661</td>
<td>0.566</td>
<td>0.456</td>
</tr>
</tbody>
</table>

Notes: NRB = negative response bias; WMT = Word Memory Test; TOMM = Test of Memory Malingering; PPV = positive predictive value; NPV = negative predictive value.
regarding positive predictive power (e.g., 0.914). TOMM appeared to have the advantage regarding negative predictive power (e.g., 0.732), which was a result of the TOMM having greater sensitivity among these data when considering the WMT with GMIP analysis. Predictive power was 100% when both tests were positive.

Discussion

This study assessed and compared the performance of the TOMM and the WMT in detecting response bias in a criminal forensic setting. To the authors’ knowledge, this is the only study presenting head to head comparisons of these two measures in a criminal forensic setting. Additionally, to the authors’ knowledge, this is merely the second study to employ a criterion-group design in comparing WMT and TOMM performance and the first to do so while considering the WMT including a GMIP profile as recommended by the test publisher (Green, 2005). Presenting statistics on the performance accuracy of PVTs in criminal forensic settings fills a gap in the current literature. It also informs clinicians working in such settings regarding test battery validity, effort, response bias, and malingering. Moreover, criterion-group methodology fulfills admissibility requirements set forth by Daubert and thus supports the suitability of these measures in a criminal forensic milieu (Bianchini et al., 2001). Given the weighty questions for which these evidentiary instruments are employed and their high utilization rates (Lees-Haley, Smith, Williams, & Dunn, 1996), performance critiques and calibration studies are imperative. Therefore, discerning the clinical accuracy and performance of these measures for use in current diagnostic criteria of malingering is necessary.

Results supported the utility of the TOMM and the WMT for identifying negative response bias in a criminal forensic setting. As suggested by Gervais and colleagues (2004), the TOMM and the WMT are neither equally sensitive nor specific. Classification rates vary greatly depending on how the two tests are used. Clinically, TOMM and WMT indices of response bias are often considered individually such that a failure on any trial is indicative of less than adequate effort and alerts the clinician in forensic contexts to consider response bias and malingering. When considered in this manner differences between TOMM and WMT indices were minimal. In theory, the classification accuracy of any measure of response bias spans a continuum of absolutely correct to absolutely incorrect. ROC analysis and associated AUCs provide a measure of this diagnostic efficiency. Diagnostic efficiency, however, is relative to context and subject to the actual incidence of negative response bias or malingering in a given sample. Without a perfect measure by which comparisons can be made, the actual incidence of response bias or malingering cannot be known, negating perfect diagnostic efficiency. Absent of perfect accuracy, performance characteristics of relative measures at the individual level are critical to clinical decisions. ROC analysis and resulting AUCs do not possess the ability to differentiate response bias case by case. Moreover, in the absence of context, expected performance and concurrent predictive power cannot be anticipated, rendering overall classification accuracy less useful to the clinician considering individual performance within specific contexts. Clinically speaking, the WMT with GMIP considered appears most proficient for ruling in intentionally produced poor effort and malingering, also known as positive predictive power. Thus, when a less than adequate performance is observed on the WMT and a GMIP profile is not produced, the clinician working in a criminal forensic arena can have a strong degree of confidence in suspecting poor effort or response bias while being alerted to the possibility of malingering. Conversely, passing scores on the WMT, particularly with a GMIP profile without corroborating observational evidence, do not necessarily communicate an absence of response bias, especially in the presence of external motivation. This issue is particularly magnified in high base rate settings (Ardolf et al., 2007) and corresponds to the WMT’s negative predictive power.

Misclassifications in the opposite direction (i.e., false positives) may be more characteristic of the WMT when GMIP is not considered, as found in previous studies (Greve et al., 2008). When incorporating the GMIP analysis for failed WMT cases, the false-positive rate of the WMT fell well below that of the TOMM (Table 4). When compared with Greve and colleagues (2008), in this sample, the WMT without consideration of a GMIP profile produced very similar characteristics. Overall, the WMT without consideration of GMIP here had slightly higher sensitivity and slightly lower specificity than that previously demonstrated.

Separate from the obvious difference in use of GMIP, a combination of differences between the studies including design characteristics (i.e., fixed vs. mixed neuropathological diagnosis), group characteristics (e.g., civil litigants vs. criminal litigants), and group perception of putative costs and benefits (e.g., monetary gain vs. loss of liberty) might affect individual performance and related outcomes. It is critical to incorporate GMIP analysis to WMT performances that fall below cutoffs for IR, DR, or CNS. Additionally, it is important to include a comparison between failed performances on the WMT for any particular case with the demonstrated functional ability presented by that case. In other words, a failed WMT when the subject does not demonstrate significant functional deficits is much more likely a true positive than a false positive, even when considering GMIP.

Employment of the GMIP significantly raised the specificity of the WMT. In this sample, the GMIP was negative (i.e., a “failing” WMT) for two individuals who otherwise produced passing scores on all other freestanding tests administered, including the TOMM. Therefore, in a minority of cases, GMIP failed to identify genuine cognitive impairment. It should be noted that these
two cases met criteria for a GMIP profile other than that the mean of their easy tests was only ~26 points greater than the mean of their difficult tests. Overall, employing GMIP in this sample resulted in the expected increase in specificity but significantly lowered sensitivity of the WMT. These data clearly reveals the overall benefit of including GMIP analysis for the WMT.

Another point that deserves discussion is that of the different ethnic composition of the +MND and −MND groups. The +MND group had significantly more Hispanic participants. Further analyses of these data were also undertaken to determine if these examinees were classified accurately. Of those of Hispanic descent in the +MND group, exactly half had produced a significantly below random performance on one or more of the freestanding PVTs administered, outside of the WMT or the TOMM. These findings support the hypothesis that these cases do genuinely meet criteria for +MND.

Used clinically, the GMIP considers performance pattern in the presence of neuropathological diagnosis, which in theory should make the GMIP, assuming proven efficiency, a worthy compliment in forensic cases where malingering exists as a standard rule out. It is necessary, however, to have a full understanding of GMIP performance relative to specific neuropathology and this highlights the importance of future investigation. That two cases in this data set were classified as −MND and failed to obtain GMIP results highlights the importance of comparing GMIP findings to real-world functioning. If the presenting diagnostic consideration and real-world functioning of the examinee is not consistent with significant cognitive compromise as presented by GMIP, then exaggeration or malingering is likely, and conversely, if there is evidence of genuine impairment but the individual produces a “near miss” on the GMIP, their performance may still be due to their neuropathology.

When compared with the prior study comparing the TOMM and the WMT and combinations thereof (Greve et al., 2008), as previously alluded to, quite different results were found in this study. Although there were a number of methodological differences between the studies, this is believed to be primarily attributable to the consideration of GMIP profiles in the current study. Whereas the WMT was discerned as having 85% sensitivity previously, here its sensitivity was only 56%. Its specificity, however, improved from 70% to 94.7%; a worthwhile tradeoff. The TOMM demonstrated a slight variance in performance as well, with sensitivity and specificity previously demonstrated at 56% and 98%, respectively, whereas here the corresponding values were 68.2% and 86.8%. These changes in the sensitivity and specificity of each test modified their corresponding classification accuracy when used together, resulting in a lower sensitivity and higher specificity when used in combination.

Given the discrepancies between the two measures illustrated here, the following question is raised: barring significantly below-chance performance, what quantity of evidence is sufficient to levy a definitive opinion of negative response bias or malingering? Such discrepancies are indicative of the inherent problems in criterion-group research of response bias and malingering (Greiffenstein, Baker, & Gola, 1994; Rogers, 2008) and illustrate the importance of utilizing multiple measures, understanding the context in which the measures are utilized, and employing criteria such as Slick and colleagues (1999).

Limitations and Future Studies

Despite utilizing current state of the art in creating known groups in this study, several limitations are worth noting. The heterogeneity of the group, while likely comparative to many clinical settings, prevents understanding the performance of these measures with specific clinical pathology and thwarts the ability to account for performance relative to neurological or psychiatric disease. Similarly, the sample was administered varying PVTs, with each PVT having unique sensitivity and specificity; as such, any given examinee may be classified by a combination of tests with different paired sensitivity and specificity values than those for any other examinee. In the same vein, one of the tests used to classify subjects was the MSVT, which is conceptually related to the WMT; this similarity between the tests may have created some over-estimation of agreement between the two tests.

Some discussion of the demographic characteristics of the groups is also warranted. Although the +MND group demonstrated significantly lower IQ than the −MND group and this is suspected to be caused by performance suppression, a higher prevalence of actual borderline intellectual ability in this group cannot be ruled out as individual case histories were not reviewed for this study. Alternately, a brief analysis demonstrated that at least 17 of the 41 individuals in this group produced a below random performance on one or more on the criterion tests, indicating that this was likely not the case in at least a sizeable subgroup of these examinees. Also, a number (n = 9) of Hispanic individuals were administered the VIP Verbal subtest. Although clinical judgment was used regarding whether or not the test would be appropriate for the individual at the time of evaluation, again since individual case histories were not examined for this study, English language proficiency of these individuals is not assured. It should be noted, however, that five of these individuals produced valid profiles on this test, whereas three produced irrelevant profiles and one produced a suppressed profile.

Another limitation, common to all criterion-group studies of response bias and malingering, is the lack of a perfect measure of response bias by which to judge TOMM and WMT performance (Mossman, Wygant, & Gervais, 2012). The utilization of other well-known PVTs for establishing criterion groups, while necessary and the best possible alternative, is prone to creating unknown error. As stated by Greve and colleagues (2008), the two aforementioned limitations will serve only to create more reserved cut scores. Ultimately, reserved cut scores and extended efforts on the part of the clinician serve to protect the examinee, clinician, and
field from the detriment of false-positive errors. In addition to the unknown error created by the lack of a perfect measure of malingered, criterion-group methodology contains inherent weaknesses such as the relegation of inferences to individuals who were detected and the latent insensitivity to cases with more subtle performance patterns (Lezak, Howieson, & Loring, 2004).

In summation, direct comparisons of TOMM and WMT performance were conducted employing subjects referred for neuropsychological examination within a criminal forensic hospital. Generally, when considered as a whole, the WMT appears to have a high false-positive error rate only when considering IR, DR, and CNS at manual cutoffs apart from GMIP analysis. This finding is mitigated significantly by the consideration of GMIP analysis, which increases the specificity of the test. The TOMM also has good sensitivity and specificity with a criminal forensic population. Overall, the TOMM and the WMT have good psychometric properties when employed with criminal forensic evaluatees, supporting their use with this population. This research further validates the importance of forensic practitioners using the full profile analysis for the WMT.

Conflict of Interest

None declared.

References


