Test of Memory Malingering and Word Memory Test: A new comparison of failure concordance rates

Manfred F. Greiffenstein\textsuperscript{a,}\textsuperscript{*}, Kevin W. Greve\textsuperscript{b,c}, Kevin J. Bianchini\textsuperscript{b,c}, W. John Baker\textsuperscript{a}

\textsuperscript{a} Psychological Systems, Inc., Royal Oak, MI, United States
\textsuperscript{b} University of New Orleans, New Orleans, LA, United States
\textsuperscript{c} Jefferson Neurobehavioral Group, Metairie, LA, United States

Accepted 31 July 2008

Abstract

Two commonly used symptom validity tests are the Test of Memory Malingering (TOMM) and Word Memory Test (WMT). After examining TOMM–WMT failure concordance rates, Green [Green, P. (2007). Making comparisons between forced-choice effort tests. In K. B. Boone (Ed.), Assessment of feigned cognitive impairment (pp. 50–77). New York: Guilford] urged widespread adoption of the WMT, arguing the TOMM is insensitive to feigned impairment. But Green (2007) used a skewed concordance method that favored WMT (one TOMM subtest vs. three WMT subtests). In the present study we compare pass/fail agreement rates with different combinations of TOMM and WMT subtests in 473 persons seeking compensation for predominately mild neurological trauma. We replicated Green (2007) using his asymmetrical method, but otherwise we found the WMT and TOMM produce comparable failure rates in samples at-risk for exaggeration with balanced comparison (three TOMM subtests vs. three WMT). Further work is necessary to compare WMT and TOMM specificities, as failure concordance designs establish reliability but are insufficient for proving validity.

© 2008 National Academy of Neuropsychology. Published by Elsevier Ltd. All rights reserved.

Keywords: Malingering; Postconcussion; Chronic pain; Validity testing; Test of Memory Malingering; Word Memory Test

Assessment of test-taking effort with symptom validity tests (SVTs) is a feature of modern neuropsychological testing (Boone, 2007; Larrabee, 2007). Neurological injury is a context where benefits can potentially be gained by feigning functional impairment. Financial incentive has proven association with outcome in remote head injury (Binder & Rohling, 1996; Rohling, Binder, & Langhinrichsen-Rohling, 1995) and chronic pain (Bianchini, Curtis, & Greve, 2006; Harris, Mulford, Solomon, van Gelder, & Young, 2005). Other justifications for SVT use include malingering base rates sufficiently high to justify validity testing in patients with ambiguous neurological injury (Greiffenstein & Baker, 2006; Greiffenstein, Baker, & Gola, 1994; Greve, Bianchini, Black, et al., 2006; Greve, Bianchini, & Doane, 2006; Larrabee, 2005; Mittenberg, Patton, Canyock, & Condit, 2002; Ord, Greve, Bianchini, & Curtis, 2007) and a professional consensus that adequate assessment of response validity is fundamental to neuropsychological assessment (Bush et al., 2005). However, sufficient justification for SVT use creates a new issue: Choice of particular SVT. There

\textsuperscript{*} Corresponding author at: ABPP-CN, 32121 Woodward Avenue, Suite 201, Royal Oak, MI 48073, United States. Tel.: +1 248 398 2200; fax: +1 248 398 2280.

E-mail address: mfgreiff@comcast.net (M.F. Greiffenstein).
are many SVTs with variations in method (forced choice, level of performance) and stimulus content (verbal, pictorial) (Bianchini, Mathias, & Greve, 2001), including Test of Memory Malingering (TOMM; Tombaugh, 1995; Tombaugh, 2002), Word Memory Test (WMT; Green, Lees-Haley, & Allen, 2002), Victoria Symptom Validity Test (Slick, Hopp, Strauss, & Spellacy, 1996), Rey Fifteen Item Test (Frederick, 2003), and many others (Boone, 2007). There are too few comparative studies to guide choice.

One basis for choosing an SVT is concordance analysis, or comparison of sensitivities. Concordance means a comparison of agreement in pass/fail rates for two or more SVTs, while sensitivity means accuracy identifying persons known to have a trait or condition (Larrabee & Berry, 2007). In many studies, Paul Green and his colleagues compared WMT and other SVT pass/fail rates in neurological samples at-risk for exaggeration (Green & Flaro, 2003; Green & Iverson, 2001; Green, Iverson, & Allen, 1999; Green et al., 2002). Paul Green summarized his concordance research in a recent book chapter. Green (2007) reported that his at-risk (for exaggeration) sample failed the WMT roughly three times more often than the TOMM; he concluded that the WMT was more sensitive to malingering and should be the preferred and possibly the only SVT.

A critical reading of Green (2007) suggests a methodological limitation. Green calculated concordance by comparing failure rates on a single TOMM subtest (Trial 2) to failure rates on any of three simple WMT subtests (Immediate and Delayed Recognition, Consistency). Providing multiple opportunities to fail on one measure, but only a single opportunity to fail another, is a biased comparison. Any statistical outcome depends on number of opportunities; aggregated malingering indicators are more powerful predictors than any single indicator (Larrabee & Berry, 2007). In essence, the SVT comparison method used by Green (2007) may have biased the results in favor of the WMT.

We revisited the issue of TOMM/WMT failure concordance with new aggregation formats. Our goal was to examine the impact of symmetrical versus asymmetrical score aggregation on concordance rates. Using a large sample of persons seeking compensation for remote neurological trauma, we examined agreement rates in three formats: (a) Green’s (2007) original asymmetrical aggregation (three WMT subtests versus one TOMM subtest), (b) a symmetrical method using all three TOMM subtests versus three WMT, and (c) reversed asymmetry potentially favoring the TOMM (all three TOMM subtests, one WMT subtest).

1. Methods

1.1. Participants

TOMM and WMT data were taken from the files of 473 persons referred for neuropsychological testing in Michigan (Greiffenstein and Baker) and Louisiana (Greve and Bianchini). All participants were seeking compensation for subjective disability (work, home, or school) blamed on remote neurological trauma (>6 months since injury) associated with either persistent cognitive or chronic pain complaints; most injuries were mild or ambiguous and not predictive of chronic disability, with more serious cases represented in the in the LA dataset. All data was archival but consecutive; referral sources included claims managers, physicians, and attorneys. The Glasgow Coma Scale (GCS), length of posttraumatic amnesia (PTA), radiographic findings (head CT/MRI/MRA), and initial neurological abnormalities were used to distinguish mild and serious initial injuries. The total sample was 90.5% comprised of persons with initial neurological trauma that was objectively mild and 9.5% comprised of quantitatively serious trauma.

The social and injury characteristics of the two samples and total sample are summarized in Table 1. The Michigan sample (n = 136) was comprised primarily of persons with persistent postconcussion syndrome (PPCS; n = 127), meaning polysymptomatic report long after minor head or neck trauma; their admission Glasgow Coma Scales (GCS) were 13–15 in the ER and PTA <30 m or no PTA at all. Nine persons sustained moderate-severe traumatic brain injury (M/S TBI) (admitting GCS was 3–12, PTA > 24 h, plus radiographic abnormalities). Using the same initial injury and duration criteria, the Louisiana sample (n = 337) was comprised of 21.7% (n = 73) patients with PPCS presentations after their mild injuries, 67.7% with chronic pain presentations after mild injury, and 10.7% with M/S TBI histories. The data of these 36 M/S TBI patients is also reported in a separate study (Greve, Ord, Curtis, Bianchini, & Brennan, 2007). The chronic pain group consisted of 228 persons claiming disabling pain and who were not claiming a head injury. Of the sample, 89% reported spine pain (neck or back) and 93% reported non-spine pain, usually involving the upper extremities (67%), less often the head, chest, or abdomen (33%). Approximately 1% had been diagnosed with fibromyalgia, myofascial pain syndrome, or complex regional pain syndrome. Half the patients (54%) had undergone surgery. Twenty-six percent had received a discectomy/fusion and 11% had a decompression/laminectomy. Seventeen
percent had undergone a non-spine surgical procedure including repair of fractures, arthroscopic surgery, and carpal tunnel release. These patients rated their pain at the time of the clinical interview at $M$ of 6.59 ($SD = 2.03$) on a 10-point scale with 0 being no pain and 10 reflecting the worst imaginable pain.

### 1.2. Procedure

The TOMM (Trial 1, Trial 2, and Retention) was administered according to the standardization manual (Tombaugh, 1995). Although the manual makes TOMM Retention optional, it was customary practice of all authors to administer the Retention trial regardless of Trial 1 and 2 scores. In both clinics the option to omit the retention trial was not exercised in the normal scope of the clinical use of this test. The WMT (Immediate Recognition, Delayed Recognition, and Consistency) was administered via laptop computer per the administration manual (Green, 2005). Our TOMM and WMT pass/fail rules were copied from Green (2007). A TOMM score <90% correct (i.e., <45/50 correct) was coded as a TOMM failure, and a WMT score of $\leq 82.5\%$ correct was defined as WMT failure; passes were scores above these cutoffs. The only departure from Green (2007) is our application of the same cutting scores to all three TOMM subtest, not just Trial 2. The point of the present study is to make fairer comparisons to the three primary WMT subtests. We incorporated Trial 1 (a) to make comparisons with WMT symmetrical and (b) subsequent research showed Trial 1 adds incremental validity (Greve, Bianchini, Black, et al., 2006; Greve, Bianchini, & Doane, 2006; Greve et al., 2007).

Although Tombaugh (1997) did not recommend a cutting score for Trial 1 there are published classification accuracy data for Trial 1 to help understand the potential magnitude of detection increases which might be expected. Greve, Bianchini, and Doane (2006) demonstrated a false positive error rate of 15% in mixed TBI at a Trial 1 cutoff of $<45$ and Greve, Bianchini, Black, et al. (2006) and Greve, Bianchini, and Doane (2006) found it to be 12% in patients alleging toxic exposure. At the same cutoff, Trial 2 and Retention had extremely low FP rates (0% in toxic exposure and less than 5% in TBI). These latter values are important because Greve et al. (2007) reported on the relative false positive error rates in the TOMM and WMT in TBI and pain using data independent of the two previously cited studies; the TOMM data did not include Trial 1. This study replicated the earlier findings, demonstrating a false positive error rate of less than 5% in TBI and less than 2% in pain in Trial 2 and Retention. When these data were re-analyzed including a Trial 1 cut score of $<45$, the test-wide FP rate was 15%. This same study demonstrated WMT test-wide FP rate of 30% TBI and 10% in pain.

### 2. Results

#### 2.1. Replicating Green’s method

The first concordance analysis copied Green’s (2007) method, which compared failure on TOMM Trial 2 with aggregated failures of WMT (any subtest failed). Table 2 summarizes the agreement rates. Overall agreement was seen
Table 2
Agreement between TOMM and WMT in 473 respondents using Green (2007) asymmetrical method

<table>
<thead>
<tr>
<th>TOMM</th>
<th>WMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michigan</td>
<td>Louisiana</td>
</tr>
<tr>
<td>Pass n (%)</td>
<td>Fail n (%)</td>
</tr>
<tr>
<td>Pass</td>
<td>56 (41.2)</td>
</tr>
<tr>
<td>Fail</td>
<td>3 (2.2)</td>
</tr>
</tbody>
</table>

Asymmetrical method: comparing failure rates on TOMM Trial 2 versus any failure on three WMT subtests; TOMM = Test of Memory Malingering; WMT = Word Memory Test; percent refers to within-sample percentage passing or failing.

Table 3
Agreement between TOMM and WMT using symmetrical test score comparisons

<table>
<thead>
<tr>
<th>TOMM</th>
<th>WMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michigan</td>
<td>Louisiana</td>
</tr>
<tr>
<td>Pass (%)</td>
<td>Fail (%)</td>
</tr>
<tr>
<td>Pass</td>
<td>43 (31.6)</td>
</tr>
<tr>
<td>Fail</td>
<td>16 (11.8)</td>
</tr>
</tbody>
</table>

Symmetrical method means all three TOMM subtests (Trials 1 and 2, Retention) compared to all three simple WMT subtests (Immediate and Relayed Recognition, Consistency).

In 341 of 473 cases (summing the joint pass/pass and fail/fail cells), or 72.1%. This is not substantially different from the 74.3% agreement report shown in Green’s (2007) Table 4.1 (p. 53). In our sample, the tests disagreed 27.9% of the time, compared to the 25.7% reported by Green (2007). Regarding the critical issue of differential sensitivity, 129 persons (27.3% of total) failed the WMT but passed the TOMM, while only three failed TOMM but passed WMT (1% of total). These findings are nearly identical to Green’s (2007) findings of 25% failing WMT (but passing TOMM) and 1% failing TOMM (but passing WMT). In conclusion, we replicated Green’s (2007) main finding, proving we are dealing with reliable phenomena across different settings.

2.2. Symmetrical concordance analysis

The first analysis replicated Green’s (2007) findings to near-parity. Hence, we can be certain the two SVTs are reliable across different laboratories. But the purpose of this paper is to evaluate whether balancing indicators affects pass/fail rates. To this end, we provided equal aggregation for TOMM scores by defining a TOMM failure as a score of <45/50 on any subtest. The definition of WMT failure was kept constant.

Table 3 summarizes the concordance rates with the symmetrical comparison. First, the overall agreement rate is 77.2%, nearly identical to the 75% reported by Green. But second, the differential sensitivity rates are more similar, with 65 persons (13.7%) failing WMT but passing TOMM, compared to 43 (9.1%) failing TOMM but passing WMT. In other words, the WMT had an incremental advantage of only 21 persons (65 – 43, or 4.4% of total sample) out of 473 cases, versus the incremental “advantage” of 126 persons (129 – 3, or 26.6%) using Green’s biased method as described above.

However, even this small 4.4% higher positive rate for WMT does not automatically constitute incremental sensitivity or validity, if one considers prospects for false positive errors. A higher positive (failure) rate with one test can only represents two possibilities: A true positive finding or a false positive error. Based on the well-founded assumption that persons with more severe brain injuries are less likely to exaggerate than persons seeking compensation for minor injury (Green & Iverson, 2001; Greiffenstein & Baker, 2006), a higher “failure” rate for the WMT than TOMM raises questions about false positive errors. To address this, we examined concordance rates for the 45 persons with M/S TBI (nine from Michigan, 36 Louisiana). Of the 45 persons with M/S TBI, TOMM and WMT agreed in 36 cases (80%). But seven M/S TBI patients (15.6%) failed WMT but passed TOMM, compared to two seriously injured patients (4.4%) who failed TOMM but passed WMT. Although malingering in this severely injured subsample cannot be ruled out,
Table 4
Agreement between all TOMM indicators and only WMT delayed recognition in 473 respondents (reversing Green’s (2007) asymmetrical method)

<table>
<thead>
<tr>
<th>TOMM</th>
<th>WMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michigan sample (N = 136)</td>
<td>Louisiana sample (N = 337)</td>
</tr>
<tr>
<td>Pass (%) Fail (%)</td>
<td>Pass (%) Fail (%)</td>
</tr>
<tr>
<td>Pass 49 (36) 9 (6.6)</td>
<td>172 (51.0) 35 (10.4)</td>
</tr>
<tr>
<td>Fail 27 (19.9) 51 (37.5)</td>
<td>42 (12.5) 88 (26.1)</td>
</tr>
</tbody>
</table>

WMT’s higher positive rate in severe brain injury is better explainable by false positive error. This will need to be re-examined with a larger sample of seriously injured persons.

2.3. Aggregation favoring TOMM

The prior analysis showed that given a balanced comparison, the TOMM produces a comparable rate of positive findings to the WMT. We next sought to demonstrate that redefining TOMM failure as failing any TOMM subtest gives an advantage over a single WMT indicator. To this end, we defined TOMM failure as any score <45/50 on Trial 1, 2, and Retention. We defined WMT failure as a score \( \leq 82.5\% \) on WMT Delayed Recognition (WMT-DR) only. WMT-DR is comparable to TOMM Trial 2 because of its position as second memory probe after two study opportunities. Table 4 shows the results. The agreement rate was 76.1%. This time, 69 (14.6%) of persons “failed” TOMM but only 44 (9.3) “passed” WMT, giving TOMM an incremental advantage of 25 persons (69 − 44, or 5.2%). Hence, slightly more test-positive results are obtained if one characterizes TOMM failure in the same way that Green (2007) characterizes WMT failure.

3. Discussion

We sought to compare the failure concordance of the TOMM and WMT under different pass/fail definitions. The main result is that the rate of positive findings (percent failing an indicator) differed dramatically depending on the indicator aggregation method. We replicated Green’s (2007) showing of a much higher WMT failure rate when any WMT subtest failure was compared to a single TOMM indicator. But under our symmetrical (and fairer) concordance method, both the WMT and TOMM produced more similar failure rates. Further analysis showed WMT failed more often than TOMM by the moderate-severe brain injury subsample. Our main conclusion is that belief in WMT superiority over the TOMM is unfounded.

Our study results do not mean that WMT or TOMM should be chosen by coin flip, nor do our findings imply equal validity. Also, we are not offering any TOMM Trial 1 cutting score for clinical use; we simply used Trial 1 for purposes of a more balanced concordance analysis, not for validity analysis. Other choices of cutting score may have given different results. A concordance analysis is just a variant of the differential prevalence design, a format that compares the rate of positive findings in the absence of external criteria for malingered cognitive deficit. The comparative validities of the WMT and TOMM have not yet been thoroughly investigated. Although at least modest incremental sensitivity (relative to all other methods) is a necessary feature of an SVT, the more important quality is strong specificity, preferably at or above 90%; this is important to avoid false positive errors (Larrabee & Berry, 2007). In this study, it should not be concluded that the slightly higher WMT failure rate (in the fair comparison condition) means incrementally better sensitivity (more false negatives for the TOMM); it could also mean higher false positive errors for the WMT (better specificity for the TOMM). Differential prevalence designs cannot prove sensitivity or specificity because there is no external (independent) criterion of simulation in order to separate true from false positives. This design can only inform specificity if the sample is genuinely impaired and without obvious incentive. At best, differential prevalence is most useful for pilot studies to identify samples at risk for exaggeration, but not sufficient in themselves for determining classification accuracy (Greve & Bianchini, 2004). For example, Gervais, Green, Allen Iii, and Iverson (2001) demonstrated that 40% of chronic benign pain patients groups fail at least one subtest of the WMT, but the mentally retarded do not.
Only a known-groups design can address incremental validity when comparing two or more SVTs (Greiffenstein, 2007). Preliminary evidence recently we partially republished here, and developed by Greve et al. (2007) showed that the published cutoffs for the WMT have a higher false positive error rate than the TOMM at its published cutoffs, using some of the same subjects studied in this paper. When fairly compared either using Receiver-Operator Characteristic analysis or comparing sensitivity at scores associated with approximately equal specificity, the two tests are nearly identical in prediction.

References


