Pergamon

Archives of Clinical Neuropsychology 19 (2004) 475–487

Archives of CLINICAL NEUROPSYCHOLOGY

A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants

Roger O. Gervais [a,*], Martin L. Rohling [b], Paul Green [a], Wendy Ford [a]

[a] *Neurobehavioural Associates, Suite 201, 17107-107 Avenue, Edmonton, Alta., Canada T5S 1G3*
[b] *University of South Alabama, Mobile, USA*

Accepted 19 May 2003

Abstract

Two-alternative forced-choice procedures have been the most widely employed for detecting incomplete effort and exaggeration of cognitive impairment. However, it cannot be assumed that different symptom validity tests (SVTs) are of equal sensitivity. In this study, 519 claimants referred for disability or personal injury related assessments were administered three SVTs, one based on digit recognition (Computerized Assessment of Response Bias, CARB), one using pictorial stimuli (Test of Memory Malingering, TOMM) and one employing verbal recognition memory (Word Memory Test, WMT). More than twice as many people failed the WMT than TOMM. CARB failure rates were intermediate between those on the other two tests. Thus, tests of recognition memory using digits, pictorial stimuli or verbal stimuli, all of which are objectively extremely easy tasks, resulted in widely different failure rates. This suggests that, while these tests may be highly specific, they vary substantially in their sensitivity to response bias.
© 2003 National Academy of Neuropsychology. Published by Elsevier Ltd. All rights reserved.

Keywords: WMT; CARB; TOMM; Effort; Disability; Malingering

The use of forced-choice symptom validity tests (SVTs) to detect exaggerated cognitive problems has become increasingly widespread and accepted in clinical neuropsychology (Franzen & Iverson, 1997, 1998; Iverson & Binder, 2000; Nies & Sweet, 1994; Reynolds, 1998; Rogers, 1997). Various formats have been utilized, including digit recognition pro-

* Corresponding author. Tel.: +1-780-486-6633; fax: +1-780-484-5552.
E-mail address: rgervais@shaw.ca (R.O. Gervais).

0887-6177/$ – see front matter © 2003 National Academy of Neuropsychology.
doi:10.1016/j.acn.2003.05.001

cedures (Allen, Conder, Green, & Cox, 1997; Hiscock & Hiscock, 1989), word recognition (Green, Allen, & Astner, 1996), and picture recognition (Tombaugh, 1996). The initial research on the forced-choice format used failure criteria based on a worse than chance paradigm. Further work in the field led to the development of clinical norms based on the performance on patients with documented brain trauma. With the establishment of brain injury referenced norms, it was possible to use statistically derived cutoff scores to determine failure rather than rely on below-chance performance (Rogers, 1997). Although this move to using cutoff scores has enhanced the clinical sensitivity of forced-choice procedures in detecting exaggeration of cognitive symptoms, it cannot be assumed that all symptom validity testing procedures will be equally sensitive in all cases. Indeed, as noted by Tombaugh (1996), patients may fail one type of symptom validity test more than another because of the relevance of the test material to their presenting complaints. The present study was designed to evaluate the sensitivity and specificity of three commonly administered forced-choice symptom validity tests in a sample of clinical cases presenting for disability or personal injury assessments.

1. Method

1.1. Participants

Archival data were compiled from the files of 519 consecutive adults referred to the first author (R.G.) for psychological or vocational assessments. The entire sample was 61% male, averaged 40 years of age (S.D. = 10.7), and had 11.5 years of education (S.D. = 2.5). Approximately 86% of claimants spoke English as a first language and all were rated as fluent in English. The majority of cases (66%) had orthopedic diagnoses involving musculoskeletal injuries to the low back, upper extremities, or neck. The next most common diagnostic class was comprised of claimants with fibromyalgia or chronic fatigue syndrome (9%). A description of the diagnostic classes and pain sites is contained in Table 1.

All claimants were involved in workers' compensation, long-term disability, or personal injury claims. Workers' Compensation Board (WCB) referrals comprised approximately 69% of the sample. All participants were assigned to one of two groups. The Psychological group consisted of 326 claimants (63% of sample) who had been assessed to determine their eligibility for long-term disability benefits, or to evaluate the extent of pain or other psychological damages in the context of personal injury litigation. In this group, there was the potential for financial gain, if they demonstrated impairment. In contrast, the Vocational group consisted of 193 claimants (37% of sample) who had been assessed to determine their suitability for vocational retraining because they could not return to their former occupation. All of these claimants had been deemed medically fit to return to work (with physical restrictions) and were at or near the end of their claims. If they demonstrated their competence on testing, there was the possibility that they could receive funding for further education or training and extended disability benefits during the retraining period. The two groups were relatively similar on most of the demographic variables, with the exception of significant differences in mean age and Verbal IQ.

Table 1
Primary diagnoses and pain sites in the study sample (%)

Diagnostic class	Total group ($N = 519$)	Psychological ($n = 326$)	Vocational ($n = 193$)
Orthopedic	65.5	67.2	62.7
FM/CFS[a]	8.7	13.8	0.0
PTSD[b]/Anxiety	7.1	11.0	0.5
Repetitive strain	4.8	0.6	11.9
Depression	2.5	2.1	0.5
Pain site			
Head/TMJ[c]	11.2	14.1	6.2
Neck	14.1	21.2	2.1
Back	27.5	29.4	24.4
Extremities	33.0	24.3	47.7
None/not applicable	11.6	7.4	18.7

[a] FM/CFS: fibromyalgia/chronic fatigue syndrome.
[b] PTSD: posttraumatic stress disorder.
[c] TMJ: temporomandibular joint.

1.2. Symptom validity tests (SVTs)

In the present study, three symptom validity tests were administered. The first was the Word Memory Test (WMT; Green et al., 1996). This is a word list learning task and we used the version that is administered via computer. The second was the Computerized Assessment of Response Bias (CARB; Allen et al., 1997; Conder, Allen, & Cox, 1992), which involves digit recognition and was administered via computer. The third was the Test of Memory Malingering (TOMM; Tombaugh, 1996), a memory for pictures task, which was administered via booklets and standardized instructions. All three tests were standardized on brain-injured patients. To reflect clinical practice, the cutoff scores for biased responding suggested by the tests' authors were used to determine the failure rates on each of the SVTs. For the WMT, we used any one of the immediate recognition (IR) or delayed recognition (DR) scores below the cutoff. In addition the person failed if they scored below the cutoff for consistency of responses between the latter two trials (Cons). For the CARB, the total score was used. Finally, for the TOMM, we used the two decision rules specified in the test manual: below-chance responding on any trial or a score lower than 45 on Trial 2 or the Retention Trial. We also observed the author's specification that the Retention Trial is optional and need be administered only if the Trial 2 score is less than 45 (Tombaugh, 1996). Thus, we used the same cutoffs, which are being widely applied by clinicians and which are reported in most other studies of these tests (Gervais, Green, Allen, & Iverson, 2001; Green & Iverson, 2001; Iverson, Green, & Gervais, 1999). A composite SVT score was generated using the same statistical procedures as described for the Overall Test Battery Mean (OTBM) composite below, with which overall performances on ability tests were summarized. z scores were generated for each dependent measure using the entire sample's means and standard deviations.

1.3. Objective performance: Overall Test Battery Mean (OTBM)

All claimants were interviewed and tested for up to two full days, depending upon their work speed and the referral type. They were administered up to 23 separate neurocognitive measures of ability, described in Appendix A. Because some claimants were either exceedingly slow, uncooperative, or had insurance coverage limitations, not all ability tests were administered to all claimants. The mean number of dependent measures per claimant was 13 (S.D. = 5.4), with a median of 15 and a mode of 17 (29% of all cases). Following the assessment, each test score was converted to a *z* score, which was relative to claimants within the sample (i.e., for each dependent variable, the entire sample's *M* and S.D. was used to calculate each person's score, so that the entire sample's mean for the dependent variable was 0 and the standard deviation was equal to 1.0). This procedure is similar to that used by Green et al. (2001), except that they generated *z* scores using external norms (i.e., means and standard deviations for normal controls that were published in a test's technical manual). Heaton et al. (2001) used a similar procedure with *T* scores (mean of 50 and a standard deviation of 10) in their investigation of the stability and course of neuropsychological deficits in schizophrenia. Relative *z* scores have the advantage over normative *z* or *T* scores of making it simpler to compare subsets of claimants within the sample. Finally, each claimant's *z* scores were averaged to generate an Overall Test Battery Mean, which was developed by Miller and Rohling (2001).

1.4. Measures of pain: acuity and location

Two primary measures of pain were available for analysis. The first of these was a self-rating of Pain Acuity during the assessment session. Ratings ranged between 0 (no pain) and 5 (extreme pain). The second measure was the number of self-rated locations within the body where pain was present at the time of assessment. The five locations that were queried by the examiner were head, neck, temporomandibular joint (TMJ), back, and extremities. The self-rating of location was binary, from 0 (no pain) to 1 (have pain). The sum of painful locations then ranged from 0 (no locations) to 5 (whole body pain). The Pearson *r* between the Pain Acuity rating and number of locations was .40 ($p < .0001$).

1.5. Self-reported memory problems: Memory Complaints Inventory (MCI)

Complaints of cognitive dysfunction are common in people with chronic pain (Iverson & McCracken, 1997). The Memory Complaints Inventory (Green & Allen, 1997) was administered to obtain an indication of the type and extent of memory problems reported by claimants at the time of assessment. The MCI is a computer administered self-report inventory of memory problems ranging from common to implausible. It contains nine scales, including: (a) General Memory Problems, (b) Numeric Information Problems, (c) Visuospatial Memory Problems, (d) Verbal Memory Problems, (e) Pain Interferes with Memory (PIM), (f) Memory Interferes with Work, (g) Impairment of Remote Memory, (h) Amnesia for Complex Behavior, and (i) Amnesia for Antisocial Behavior. Symptoms of the Pain Interferes with Memory scale were endorsed by 84% of the total sample, and 93% of the Psychological group. Eighty percent of the total sample and 87% of the Psychological group reported that memory problems interfered with their ability to work.

2. Results

2.1. Failure rates on SVTs

Of the 519 claimants in the sample, 35% failed one or more of the SVTs. Below-chance scores, as defined by the binomial distribution, were rare in all groups. The TOMM manual indicates that scores of less than 18 are in the below-chance range at the 95% confidence level. The equivalent cutoff for below-chance performance on the WMT is 37.5%. Below-chance performance on the CARB total test is less than 42%, and 35% for each block. Only four TOMM protocols out of 519 (i.e., 0.07% of all cases) contained one or more scores in the below-chance range on Trials 1 or 2 or the Retention Trial and all such cases scored below 45 on Trial 2. There were nine cases with worse than chance CARB scores and six with worse than chance WMT scores. The vast majority of failures on all three SVTs were attributable to scoring below the empirically established cutoffs but not at a below-chance level.

The rates of failure varied widely from one SVT to another. The WMT failure rate was 32%, compared to 17% for the CARB and 11% for the TOMM. The rates of SVT failure also varied substantially depending upon assessment type. Whereas 43% of the Psychological group failed the WMT, the failure rate was only 12% in the Vocational group, $\chi^2(1) = 52.2$, $p < .0005$. On the CARB, 25% of the Psychological group failed compared with 4% of the Vocational group, $\chi^2(1) = 37.9$, $p < .0005$. Finally, on the TOMM, 17% of the Psychological group failed compared to 1% of the Vocational group, $\chi^2(1) = 30.4$, $p < .0005$. The failure rates for each SVT are detailed in Table 2.

These data suggest that the WMT, CARB, and TOMM are not equally sensitive to response bias or suboptimal effort. The author-recommended decision rules (below-chance on any trial, <45 on Trial 2 or Retention Trial) for the TOMM misclassified 69% of the claimants who produced inadequate effort on at least one other test. However, scores below 45 on the TOMM were highly specific, correctly identifying suboptimal effort in nearly 100% of the cases (where failure on the WMT and/or the CARB was used as the external criterion). In fact, in those who failed the TOMM, all but one case failed the WMT. Further analysis of the TOMM data revealed that 53% of claimants scoring between 45 and 49 on Trial 2 ($n = 73$) fell below the cutoffs for biased responding on the WMT and 23% were below the

Table 2
Rate of SVT failure by assessment type

SVT	Total sample ($N = 519$) (%)	Psychological ($n = 326$) (%)	Vocational ($n = 193$) (%)
WMT	32	43	12
CARB	17	25	4
TOMM	11	17	1

cutoff on the CARB. These cases produced a mean WMT-DR score of 75.8 (S.D. $= 12.3$, $n = 39$) and a mean CARB score of 87.8 (S.D. $= 16.3$). Both of these mean scores are below the failure cutoffs recommended by the authors of these two instruments. This suggests that, even when TOMM scores are in the passing range of 45–49, there is a greater than 50% chance of missing response bias as defined by these other two validated SVTs. Of the 390 claimants who obtained a perfect score of 50 on the TOMM, 18% failed the WMT, and 8% failed the CARB. Finally, in nearly 21% of cases producing perfect scores on the TOMM, evidence of response bias was detected by one of the other two SVT measures.

It might be argued that the WMT is overly sensitive and thereby prone to false positive findings of response bias. This concern was addressed, firstly, by examining the CARB and TOMM scores of the claimants who scored below the cutoff on the WMT. This subgroup ($n = 164$) obtained a mean CARB score of 83.9 (S.D. $= 19.1$) and a mean score on Trial 2 of the TOMM of 43.9 (S.D. $= 9.2$). Thus, as a group, the claimants failing the WMT also scored below the cutoffs for biased responding on the CARB and TOMM. In those cases who passed both the TOMM and the WMT, the mean score on a third independent SVT (CARB) was 98.2% correct (S.D. $= 4.0$, $n = 354$), suggesting good effort. In contrast, for those who passed the TOMM but failed the WMT, the mean CARB score was substantially lower at 89.5% correct (S.D. $= 15.0$, $n = 109$), a borderline score for biased responding on the CARB. These findings support the view that failures on the WMT are not the result of excessive sensitivity leading to false positive determinations of response bias.

The question of false positives on the WMT can also be examined by reviewing the cognitive test performance of claimants who scored above or below the cutoffs on the WMT and TOMM. Claimants who scored below the cutoff on the WMT ($n = 74$), but who passed the CARB and the TOMM, scored an average of -0.54 on the OTBM (i.e., 0.54 S.D. below the overall sample mean). In contrast, the OTBM mean score was 0.24 in the claimants who passed all three SVTs ($n = 340$, $F(1, 412) = 109.5$, $p < .0001$). Thus, compared with claimants passing the WMT, there is a relative deficit of 0.76 standard deviations in the mean score from 13 different tests of ability in those who failed the WMT but who passed the CARB and TOMM. The magnitude of the OTBM deficit observed in claimants who failed only the WMT argues strongly against any interpretation of isolated WMT failure as being indicative of only a subtle lapse of effort that has no significant influence upon their other assessment results. When all claimants who scored below the cutoff on the WMT were selected, regardless of failure on any of the other SVTs, this resulted in a mean score of -0.82 on the OTBM (or a 1.06 S.D. deficit in the OTBM vs. those passing all SVTs). Claimants who failed only the CARB obtained an OTBM of -0.06, but those who failed the CARB in conjunction with any other SVT scored obtained an OTBM of -0.91. Claimants who failed the TOMM in addition to one or more of the other SVTs obtained an OTBM of -1.23. (Note: There was only one case who failed TOMM but passed the WMT.) Cases failing all three SVTs demonstrated the worst cognitive test performance with an OTBM score of -1.39. The effect of failing various combinations of SVTs upon the OTBM is detailed in Table 3.

The effect of failure on the various SVTs can also be seen upon the individual cognitive test scores, with claimants failing the TOMM producing, on average, the lowest scores. For

Table 3
OTBM and mean Pain Acuity rating in groups passing or failing the three different symptom validity tests

SVT results				Test score	
WMT	CARB	TOMM	*N*	Mean OTBM	Mean Pain
Pass	Pass	Pass	340	0.24	−0.30
Pass	Fail	Pass	14	−0.06	−0.09
Fail	Pass	Pass	74	−0.54	0.14
Fail	Fail	Pass	35	−0.74	0.57
Fail	Pass	Fail	17	−0.96	0.39
Fail	Fail	Fail	38	−1.39	0.67

Note. Only one case failed the TOMM without failing either of the other SVTs. No meaningful group OTBM or Pain score could be calculated.

example, the mean California Verbal Learning Test (CVLT; Delis, Kramer, Kaplan, & Ober, 1987). Short Delay Free Recall score was 11.1 (S.D. = 3.0) in the claimants who passed all three SVTs. It was 8.7 (S.D. = 3.4) in those who failed the WMT but passed the TOMM and it was 6.7 (S.D. = 3.0) in those who failed both the WMT and TOMM. A similar pattern was observed in all other cognitive measures, including IQ and olfactory identification. These findings demonstrate a clear association between WMT failure and poorer cognitive test performance, irrespective of whether TOMM is passed or failed.

It might also be argued that the three SVTs vary in their level of difficulty, and that claimants of lower cognitive ability would be more likely to fail the WMT than the TOMM, leading to the discrepancy in failure rates noted in this study. To address this question we correlated years of education (an objective estimate of ability not susceptible to effort) with the score obtained on each SVT. Weak, nonsignificant correlations were found for the CARB ($r = .048$) and the TOMM ($r = .063$). A weak, but significant correlation was observed between years of education and the WMT-DR ($r = .104$, $p < .05$, $N = 519$), but years of education would account for a negligible amount of the variance in WMT scores. Claimants who passed all SVTs were not significantly different in years of education from those who failed one or more SVTs ($t(517) = 1.459$, $p =$ ns). Limited educational attainment (and by extension, cognitive ability), therefore, could not account for the SVT failure observed in this study. Further support for this conclusion is drawn from Green and Flaro (in press) who documented WMT-IR and WMT-DR scores of 90.9% (S.D. = 15.7) and 87.5% (S.D. = 25), respectively, in a sample of children of very low intelligence ($n = 13$, VIQ = 64, S.D. = 4.5) age 11.7 years (S.D. = 2.7).

2.2. Pain reports and SVT scores

The mean 0–5 Pain Acuity rating for the entire sample was 3.1 (S.D. = 1.4). The mean Pain Acuity rating of claimants who demonstrated suboptimal performance on any one of the SVTs (35% of the sample) was 3.62 (S.D. = 1.18). This was significantly higher than the mean pain rating of 2.80 in those who passed all three SVTs (S.D. = 1.36). There was also a significant difference between pain ratings from the Psychological group ($M = 3.31$, S.D. = 1.31) and

Table 4
Pain Acuity ratings for claimants who passed all SVTs (Genuine) compared to claimants who failed any one SVT (Exaggerating) by Vocational and Psychological groups

	Genuine ($n = 244$)		Exaggerating ($n = 143$)		
	Vocational ($n = 94$)	Psychological ($n = 150$)	Vocational ($n = 13$)	Psychological ($n = 130$)	F or t
Pain Acuity rating					
M	2.48	3.00	3.15	3.66	16.0
S.D.	1.28	1.38	1.63	1.12	<.0001

the Vocational group ($M = 2.56$, S.D. $= 1.33$, $t(385) = 5.00$, $p < .0005$), with higher Pain Acuity ratings generated from claimants who failed any one of the three SVTs. Each groups' Pain Acuity ratings are shown in Table 4, divided by validity status (i.e., passing or failing any one SVT).

This suggests that claimants who scored below the cutoffs on one or more of the SVTs might have exaggerated their reports of pain, as suggested by other investigators (Schmand et al., 1998). Following the definitions of Rohling, Allen, and Green (2002), claimants who failed any SVT were assigned to an Exaggerating group ($n = 179$). Claimants who passed all SVTs were termed Genuine ($n = 340$). Exaggerating claimants were twice as likely as Genuine claimants to rate their pain at the time of testing as being at the high end of the scale (i.e., 4 or 5). Conversely, Genuine claimants were 1.8 times as likely as Exaggerating claimants to rate their pain at the low end of the scale (i.e., 0 or 1), $\chi^2(387) = 47.8$, $p < .0001$.

An analysis was undertaken to evaluate the effect of self-reported pain on the OTBM by contrasting the cognitive performance of claimants in both the Exaggerating and Genuine groups who rated their Pain Acuity at a high level. The mean Pain Acuity ratings of the Genuine claimants ($M = 4.31$, S.D. $= 0.46$, $n = 80$) were not significantly different from those of the Exaggerating group ($M = 4.33$, S.D. $= 0.47$, $n = 93$). However, the claimants in the Exaggerating group obtained a significantly lower OTBM score o f -1.54, compared to 0.14 for the Genuine claimants ($p < .0001$). These results indicate that claimants' failure on any SVT was unrelated to the immediate experience of pain during the assessment. For example, 46% of the claimants who rated their pain as being severe were capable of obtaining scores above the cutoff for suboptimal effort on each of the three SVTs.

A second analysis was conducted using two subgroups drawn from the 340 Genuine claimants. The two groups were relatively similar on most of the demographic variables, with the exception of significant differences in mean age and Verbal IQ. The subgroups were chosen to increase the statistical power to detect possible differences in neurocognitive performance associated with self-reports of pain during testing. The Pain Acuity variable was collapsed into a Low Pain subgroup ($n = 41$), whose ratings were either 0 or 1, and a High Pain subgroup ($n = 80$), whose acuity ratings were either 4 or 5. Mean Pain Acuity was markedly different between the two subgroups, $t(119) = -40.6$, $p < .0001$. However, there were no significant differences between the two groups either on the mean SVT z score or on the OTBM, as noted

Table 5
Demographic and performance comparisons between Low Pain and High Pain claimants who passed all SVTs (Genuine)

	Pain rating status					
	Low rating ($n = 41$)		High rating ($n = 80$)			
	M	S.D.	*M*	S.D.	*g* (ES)	*p*
Objective performance						
Mean SVT	0.43	0.08	0.40	0.12	−0.03	.1103
OTBM	0.38	0.54	0.14	0.68	−0.38	.0596
Demographic comparisons						
Pain rating (0–5)	0.59	0.50	4.31	0.47	−7.75	<.0001
Age in years	36.8	10.0	40.0	10.2	0.32	.0317
Years of education	11.9	2.9	11.4	2.5	0.19	.3466
WAIS-R/III or MAB VIQ[a]	101.3	12.3	93.0	13.0	0.65	.0029
WRAT-III reading subtest	44.4	5.8	41.8	6.0	0.44	.1537
English first language (% total)	95.1	–	93.6	–	–	.7320
Male (% total)	63.4	–	62.5	–	–	.9215

[a] Fifty percent of the claimants were administered the WAIS-R (Wechsler, 1981) or the WAIS-III (Wechsler, 1997), the remainder were administered the Multidimensional Aptitude Battery-II (Jackson, 1998).

in Table 5. This analysis indicated that Pain Acuity at the time of assessment was not likely to have caused failure on any one of the three SVTs or compromised claimants' performance on the cognitive tests.

As a final analysis, we attempted to increase the sensitivity of the TOMM by establishing a cutting score for Trial 1, thus providing a third criterion for TOMM failure and matching the number of failure criteria available for the WMT. The TOMM manual (Tombaugh, 1996) contains details of the various clinical samples used in the standardization process. None of the Traumatic Brain Injury (TBI) sample ($n = 45$) in the manual scored below 34 on Trial 1. A cutoff score of 33 or less on Trial 1 might, therefore, also be used to determine poor effort. Although this approach is not a standard interpretive procedure for the TOMM, we reanalyzed the data to include as failures all cases who scored 33 or less on Trial 1. Using this method raised the TOMM failure rate marginally to 58 of 519 claimants (11.2%) compared to the 56 (10.8%) who failed under the conventional criteria. This enhanced failure rate was still approximately half of that produced by the WMT-IR or WMT-DR subtests used individually, and emphasizes the relative insensitivity of the TOMM compared to the WMT.

3. Discussion

The main finding of this study was that there were markedly discrepant failure rates across three different symptom validity tests in the same sample of disability claimants. This means that clinicians employing any one of these tests alone would arrive at very different conclusions about the rate of cognitive exaggeration in a patient sample, depending on which test they employed. In each subgroup, at least twice as many cases failed the WMT compared with the TOMM. The failure rates on the CARB were intermediate between those of the WMT and the TOMM. It is unlikely that the discrepancies can be explained by differences in the rates of false positive classifications between tests (i.e., wrongly concluding incomplete effort when a full effort has been made). The validation research described in the test manuals and in published studies suggests that these tests are of approximately equal difficulty and that claimants with genuine cognitive impairment, who are making an effort, should easily pass them all. It is more likely that the tests differ in their sensitivity to exaggeration and that the TOMM has the lowest failure rate because of its relative insensitivity to cognitive exaggeration (i.e., a high rate of false negative classifications).

Although the findings of this study suggest that the TOMM is largely redundant when used in combination with the WMT and the CARB, it may be useful as a measure of the extent of poor effort. There were lower cognitive test scores in the claimants who failed the TOMM (and failed the WMT), compared to those who passed the TOMM (but failed the WMT) (Table 3). Thus, TOMM failure is associated with more extreme response bias. In addition, the mean WMT score of claimants who failed both the TOMM and the WMT was lower than the score of those who passed the TOMM and failed the WMT. When found in conjunction with WMT failure, poor scores on the TOMM provide further confirmation of the suboptimal effort detected by the WMT.

In this study, isolated CARB failure had a minimal effect on the OTBM. This suggests that such occurrences may not be particularly significant with respect to the overall validity of the test data obtained in the course of the assessment. However, CARB failure is still indicative of poor effort and should not be ignored in interpreting specific test findings or the clinical presentation as a whole. Further research is needed to better understand the effects of CARB failure on cognitive test performance in non-head injury claimants.

The significantly lower SVT failure rate in the Vocational group compared to the Psychological group (WMT: 12% vs. 43%) probably reflects a fundamental difference in motivation. Whereas both groups have a clear financial incentive, the goal in the Vocational group (qualifying for academic or retraining funding) is achieved by demonstrating one's best performance on the various assessment procedures. On the contrary, in the Psychological group, the financial goal (disability benefits) is achieved by convincingly demonstrating substantial impairment and inability to work. Some disability claimants in the Psychological group may have assumed that the more impairment they displayed, the greater their probability of receiving compensation. There were also some claimants whose motivation for vocational rehabilitation was questionable and who disputed the determination that they were fit to return to work. These cases probably accounted for the majority of the SVT failures in the Vocational group. Such failures were probably indicative of a lack of engagement in, or frank resistance toward, the vocational rehabilitation process and an attempt to underscore their perception of continued disability and unsuitability for retraining.

Many neuropsychologists choose to employ only one symptom validity test in clinical assessments. Regardless of the test chosen, the results of this study suggest that this is a risky practice, especially if it is based on an untested assumption that the instrument employed is equivalent to other SVTs. Depending on the test and assessment groups considered, the rate of incomplete effort in the present study would be estimated to be as high as 43%. These rates of SVT failure are comparable to the rates of 53% on one test, and 27% on two other tests reported by Vickery, Ranseen, Cooley, and Berry (1999). Similarly, Gervais, Russell, et al. (2001) found that 35% of a sample of fibromyalgia patients scored below the cutoffs for biased responding on the CARB and/or WMT.

In this study, we used the conventional definition of WMT failure (any one of IR, DR, or Cons below the cutoff score). This may appear to have led to a bias in favor of the WMT. However, had we defined WMT failure only based on the WMT-DR score, the WMT failure rate in the present study would still have been substantially higher than the failure rates on the CARB and the TOMM. Using only WMT-DR, 23% of the total sample and 32% of the Psychological group would have failed the WMT. If failure had been defined by WMT-IR alone, 21% of the total sample and 28% of the Psychological group would have scored below the cutoff. Using only the WMT consistency measure would have produced failure rates of 30 and 41% in the total sample and Psychological group, respectively. Thus, even when employing substantially more conservative failure criteria for the WMT, this SVT still emerged as a more sensitive measure of suboptimal effort and response bias than either the CARB or the TOMM.

We also attempted to enhance the sensitivity of the TOMM by establishing an empirically derived cutoff for Trial 1 (<34) based upon the standardization data for the Traumatic Brain Injury sample documented in the test manual. Two additional claimants failed the TOMM using this nonconventional approach, but this did not alter the overall rate of TOMM failure in the sample, which remained at 11%, half that of either the WMT-IR or WMT-DR subtests used alone. This further supports the conclusion that the WMT is a more sensitive measure of response bias than the TOMM.

It might also be argued that the placement of the SVTs in the test battery could influence performance. Guilmette, Whelihan, Hart, Sparadeo, and Buongiorno (1996), using an abbreviated version of the Hiscock Forced-choice Procedure, found more failures when it was placed at the beginning of the test battery. In the present sample, the TOMM was typically administered in the morning as one of the first formal assessment procedures. This was followed by the CARB, usually in the in the later morning, and the WMT in the early to mid afternoon. While there is the possibility that the order of administration might have influenced the SVT results, the work of Guilmette et al. (1996) suggests that this would have increased, not decreased the TOMM failure rate. Fatigue, pain, or emotional distress over the course of the assessment may also have influenced the rate of SVT failure in the present study. This question was addressed by Gervais, Green, et al. (2001) who found that these factors could not explain the high rate of SVT failure in a sample of chronic pain patients.

There is a need for further objective studies of the differential sensitivity of various SVTs to exaggeration of cognitive difficulties. Several SVTs might be shown to be equally insensitive to genuine impairment (i.e., objectively very easy), but this would not mean that one could assume that the tests are also equivalent in their sensitivity to cognitive exaggeration. Furthermore, even if two tests were of equivalent sensitivity, it is likely that many exaggerating claimants would pass one test and fail the other, owing to the intrinsic unreliability of their test performance. The findings of this study suggest that it is advisable to use multiple measures of effort in all clinical assessments. It is also important that the selection of such procedures be based on demonstrated sensitivity to poor effort.

Appendix A

California Verbal Learning Test (CVLT; n = 6): Trial 1, Trial 5, List B, Short Delayed Free Recall, Long Delayed Free Recall, and Recognition.
Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtis, 1993, *n = 2)*: Categories Achieved and Perseverative Responses.
Wide Range Achievement Test-III (WRAT-III; Wilkinson, 1993, *n = 2)*: Spelling and Arithmetic.
Wechsler Adult Intelligence Scale-Revised/III, or Multidimensional Aptitude Battery (MAB-II; n = 2): Verbal IQ and Performance IQ.
General Aptitude Test Battery (GATB; n = 6): S, P, Q, K, F, and M.
CogniSyst Story Recall Test (CSRT; Green & Allen, 1995, *n = 1)*: Immediate Recall.
Word Memory Test Memory subtests (WMT; n = 4): Multiple Choice, Paired Associates, Delayed Recall, and Long Delayed Recall.

References

Allen, L. M., Conder, R. L., Green, P., & Cox, D. R. (1997). *CARB 97 manual for the Computerized Assessment of Response Bias*. Durham, NC: CogniSyst.
Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test (CVLT) manual*. San Antonio, TX: The Psychological Corporation.
Franzen, M. D., & Iverson, G. I. (1997). The detection of biased responding in neuropsychological assessment. In A. M. Horton (Ed.), *The neuropsychological handbook: Behavioral and clinical perspectives* (2nd ed., pp. 393–421). New York: Springer.
Franzen, M. D., & Iverson, G. I. (1998). Detecting negative response bias and diagnosing malingering: The dissimulation exam. In P. J. Snyder & P. D. Nussbaum (Eds.), *Clinical neuropsychology: A pocket handbook for assessment* (pp. 88–101). Washington, DC: American Psychological Association.
General Aptitude Test Battery (GATB). (1983). *Administration and scoring manual*. Scarborough, Ont.: Nelson.
Gervais, R. O., Green, P., Allen, L. M., & Iverson, G. L. (2001). Effects of coaching on symptom validity testing in chronic pain patients presenting for disability assessments. *Journal of Forensic Neuropsychology*, *2*, 1–19.
Gervais, R. O., Russell, A. S., Green, P., Allen, L. M., Ferrari, R., & Pieschl, S. D. (2001). Effort testing in patients with fibromyalgia and disability incentives. *Journal of Rheumatology*, *28*, 1892–1899.
Green, P., & Allen, L. M. (1995). *CogniSyst Story Recall Test*. Durham, NC: CogniSyst.
Green, P., & Allen, L. (1997). *Memory Complaints Inventory.* Durham, NC: CogniSyst.
Green, P., Allen, L. M., & Astner, K. (1996). *The Word Memory Test: A user's guide to the oral and computer-administered forms, US version 1.1*. Durham, NC: CogniSyst.
Green, P., & Flaro, L. (in press). Word Memory Test performance in children. *Child Neuropsychology*.
Green, P., & Iverson, G. L. (2001). Effects of injury severity and cognitive exaggeration on olfactory deficits in head injury compensation claims. *NeuroRehabilitation*, *16*, 237–243.
Green, P., Rohling, M. L., Lees-Haley, P., & Allen, L. M. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury*, *15*, 1045–1060.
Guilmette, T. J., Whelihan, W. M., Hart, K. J., Sparadeo, F. R., & Buongiorno, G. (1996). Order effects in the administration of a forced-choice procedure for detection of malingering in disability claimants' evaluations. *Perceptual and Motor Skills*, *83*, 1007–1016.
Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtis, G. (1993). *Wisconsin Card Sorting Test manual: Revised and expanded*. Odessa, FL: Psychological Assessment Resources Inc.

Heaton, R. K., Gladsjo, J. A., Palmer, B. W., Kuck, J., Marcotte, T. D., & Jeste, D. V. (2001). Stability and course of neuropsychological deficits in schizophrenia. *Archives of General Psychiatry*, *58*, 24–32.

Hiscock, M., & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology*, *11*, 967–974.

Iverson, G., & Binder, L. M. (2000). Detecting exaggeration and malingering in neuropsychological assessment. *Journal of Head Trauma Rehabilitation*, *15*(2), 829–858.

Iverson, G., Green, P., & Gervais, R. O. (1999). Using the Word Memory Test to detect biased responding in head injury litigation. *The Journal of Cognitive Rehabilitation*, *17*, 4–8.

Iverson, G. L., & McCracken, L. M. (1997). Postconcussive symptoms in persons with chronic pain. *Brain Injury*, *11*, 783–790.

Jackson, D. N. (1998). *Multidimensional Aptitude Battery-II manual*. Port Huron, MI: Sigma Assessment Systems, Inc.

Miller, L. S., & Rohling, M. L. (2001). A statistical interpretive method for neuropsychological data. *Neuropsychology Review*, *11*, 143–169.

Nies, K. J., & Sweet, J. L. (1994). Neuropsychological assessment and malingering: A critical review of past and present strategies. *Archives of Clinical Neuropsychology*, *9*, 501–552.

Reynolds, C. R. (1998). *Detection of malingering in head injury litigation*. New York: Plenum Press.

Rogers, R. (1997). *Clinical assessment of malingering and deception* (2nd ed.). New York: Guilford.

Rohling, M. L., Allen, L. M., & Green, P. (2002). Who is exaggerating cognitive impairment and who is not? *CNS Spectrums*, *7*, 387–395.

Schmand, B., Lindeboom, J., Schagen, S., Heijt, R., Koene, T., & Hamburger, H. L. (1998). Cognitive complaints in patients after whiplash injury: The impact of malingering. *Journal of Neurology, Neurosurgery, and Psychiatry*, *64*, 339–343.

Tombaugh, T. N. (1996). *Test of Memory Malingering*. Toronto, Ont.: MultiHealth Systems.

Vickery, C. D., Ranseen, J. D., Cooley, A. T., & Berry, D. T. R. (1999). Use of malingering tests in the forensic evaluation of pain patients [Abstract]. *Archives of Clinical Neuropsychology*, *14*, 102.

Wechsler, D. (1981). *WAIS-R manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *WAIS-III manual*. San Antonio, TX: The Psychological Corporation.

Wilkinson, G. S. (1993). *Wide Range Achievement Test, 3rd edition: Administration manual*. Wilmington, DE: Wide Range Inc.