Why clinicians often disagree about the validity of test results

Paul Green*
Neurobehavioural Associates, Edmonton, Alberta, Canada

Examining the validity of test results using specialised methods is still a relatively new venture and many different approaches are taken to the same task. This paper discusses some of the reasons why discrepant results and differing conclusions may be arrived at by clinicians or researchers, depending on their theoretical and practical choices. These choices include whether to test for effort, what methods to use, how to employ effort tests, what failure criteria to apply and how to interpret individual results. Equally important is the decision about whether or not to employ effort testing to remove error from data in group research studies. No concensus has yet been reached on the need for systematic effort testing in group studies but there are indications that it should be a serious consideration because controlling for invalid data can lead to altered conclusions.

1. Introduction

Neuropsychological test results have many clinical and research applications. However, these tests require effort and unreliable, inconsistent and invalid results are produced by some patients, especially when financial compensation is available for disability. In the past, it was common practice to make a statement in a clinical report such as "The patient appeared to be putting forth a full effort" and to assume that the test results were valid without any objective evidence but this is no longer acceptable [1]. A great deal of research has now been done on objective methods for evaluating whether test results are valid or not, the major emphasis being upon the use of symptom validity tests, combined with an evaluation of the consistency between various types of evidence [2].

Even five years ago, most neuropsychologists were not routinely using symptom validity tests or other objective methods to determine whether test results were valid or not and today some still rely on clinical judgement alone. The need for systematic attention to the validity of test results was prompted by studies showing high rates of exaggeration of cognitive impairment in certain populations, such as people with mild head injuries claiming compensation [3]. More recently, similar results have been found with other patient groups, including people with chronic fatigue syndrome [4], chronic pain after whiplash injuries [5], fibromyalgia [6] and various other diagnoses [7], making it seem that the presence of financial incentives for disability is a critical factor in determining exaggeration, rather than any particular diagnosis. Non-financial reasons for putting forth incomplete effort on testing have also begun to be investigated, as illustrated by a recent report about children who freely admitted that they chose to fail effort tests and that they produced invalid results on other tests [8].

Once psychologists began to test for validity, many methods were employed for this purpose, including various symptom validity tests, analysis of test-retest and internal consistency, formulae based on existing tests and judgment of whether test results make diagnostic and biological sense [1,2]. However, in any clinical assessment, there are many complicating factors and the interpretation of symptom validity test results is no exception. With the best intentions, two psychologists or neuropsychologists may differ in their interpretation of whether test results are valid or not. These differences of opinion could arise from many sources, including (1) the methods used to examine the validity of test results (2) the criteria used to determine exaggeration and how discrepancies between effort test results are understood (3) past clinical experience of cases who exaggerated symptoms and knowledge of symptom presentation in patients with a given disorder (5) access to full information about personal circumstances affecting motiva-

*Address for correspondence: Paul Green, Ph.D., Neurobehavioural Associates, 201 17107-107 Ave., Edmonton, Alberta, Canada T5S 1G3. Tel.: +1 780 484 5550; E-mail: paulgreen@shaw.ca.

NeuroRehabilitation 16 (2001) 231–236
ISSN 1053-8135 / $8.00

tion and (6) assumptions about how other factors, such as pain, headaches, depression, fatigue, hypomania and drug use can influence test performance.

A major source of disagreement about the validity of test results arises from the differences in the methods used to measure effort (Table 1). We are all familiar with the fact that two emotional self-rating scales or two visual memory tests may appear superficially similar and yet reveal different results. Similarly, alternative methods for identifying exaggeration do not have equivalent levels of sensitivity to exaggeration of impairment and we have the added problem that individual patients may perform inconsistently. For example, the Computerized Assessment of Response Bias [7,9] and the Word Memory Test (WMT) [10] were applied to patients with various levels of head injury severity in two studies [11,12]. Both studies showed a very important paradoxical effect, which is seen only with effort tests and which will almost never be seen with ability tests. That is, those with the most severe brain injuries scored significantly higher on both the CARB and the WMT than those with the least severe head injuries.

The paradox of better performance in those with the greatest impairment is most easily explained by the fact that these tests mainly measure effort and not ability and that exaggeration of cognitive impairment is more frequent in compensation claimants with mild head injuries than in those with severe brain injuries. In the latter studies, agreement between CARB and WMT was high (approximately 87%) in classifying people's performance as valid or invalid but this means that there were 13% of cases in which the two methods disagreed. If only one of these tests were used by one psychologist and the other test were used by another, there would be disagreement with regard to the validity of test results in 13% of cases. This is significant in a sample in which only about 28% of cases were exaggerating their cognitive impairment, based on failing either one of the effort measures. On the other hand, it is important not to overlook the fact that CARB and WMT are equally easy tests to pass for patients with severe brain injuries [13]. Neither test measures ability and so we need to ask why should a person pass one effort test but fail an equally easy effort test? This will almost never happen in someone who is making a genuine effort because such people almost never fail either the CARB or the WMT. Thus, a person who fails one of two objectively easy tests but who passes the other is performing very inconsistently, which indicates that the person's test results are unreliable. If effort tests do not measure ability, then any variation in scores across different effort tests or from test to retest is due to fluctuations in effort and it represents error variance. It is widely recognised that a high variability in test scores is an important feature of the performance of exaggerators. Hence, rather than dismissing one effort test because it produces different results from another test, we need to recognise that differences in scores between two equally easy effort tests are useful. They inform us about the error present in test results and, hence, they assist us in determining the validity of the results. It is advisable to use more than one effort test [2] and to recognise and capitalise on the fact that the psychometric properties of effort tests are not the same as those of ability tests.

The type of symptom validity test that is chosen makes a difference. For example, although both tests have been shown to be passed easily by patients with severe brain injuries, the Test of Memory Malingering (TOMM) [14], a visual-spatial recognition memory test, produces much lower failure rates than the WMT, a verbal recognition memory test. This is another reflection of the abnormally high error varaince in exaggerators' test data (assuming that we regard fluctuations due to effort as representing error, when what we really wish to measure is ability). In one study [15], there was more than a 100% excess of failures on the WMT versus the TOMM and the same has been observed in several independent centres [16]. There would be a high level of disagreement between two psychologists if one tested patients with the WMT and another tested the same patients independently with the TOMM. Similarly, varying levels of disagreement would be expected between psychologists using the WMT or the TOMM versus other methods for determining effort, such as formulae derived from the Category Test [17]. Fortunately, there are ways of determining which method is right in cases of disagreement. For example, the subgroups passing TOMM and failing WMT in the two previously mentioned studies [15,16] did not have diagnoses implying any cerebral dysfunction but they scored significantly lower on memory tests than patients with severe neurological diseases, meaning that they were probably exaggerators. The reverse pattern (failing TOMM and passing the WMT) occurred in less than one percent of cases and so further analysis of this subgroup was impossible.

The excess of WMT versus TOMM failures was explained as resulting from a generalized bias in most patients towards complaining of verbal memory problems to a much greater degree than other types of memory problems [16]. A verbal memory test might be more

Table 1
Summary of symptom validity tests mentioned in this paper

Name	Abbreviation	First author	Stimulus type	Task type
Amsterdam Short Term Memory Test	ASTM	Schmand, B.	Word Clusters	Recognition Memory
Computerised Assessment of Response Bias	CARB	Allen, L.M.	Digit Strings	Recognition Memory
Portland Digit Recognition Test	PDRT	Binder, L.	Digit Strings	Recognition Memory
Test of Memory Malingering	TOMM	Tombaugh, T.	Pictures of objects	Recognition Memory
Victoria Symptom Validity Test	VSVT	Slick, D.	Digit Strings	Recognition Memory
Word Memory Test	WMT	Green, P.	Paired Associate Word List	Recognition Memory

relevant to the person's subjective complaints than either a visual memory test or a numerical test and, for this reason, it might be more likely to be affected by exaggeration [18]. In a discrepant case, both the TOMM and the WMT would, in a sense, be right in that the person might choose to fail the WMT because they want to appear to have impaired verbal memory (exaggeration present) but they might choose to take a different approach to the TOMM and pass it (no exaggeration). In this case, exaggeration would be present but in a selective way and the person's other test results would be of doubtful validity. Similarly, exaggeration will affect some ability tests (such as a word list learning test) more than others (such as a naming test), presumably dependent upon whether or not the person perceives the test as being relevant to their complaints, a process discussed by Tombaugh [14].

Whatever the reasons for different rates of failure on different effort tests, such differences do occur and a major source of disagreement about the validity of test results is the method chosen to evaluate effort. For this reason, more comparative studies are needed to study differences between alternative ways of evaluating effort, by applying more than one method to the same groups of patients and comparing failure rates. In performing such studies, it will be helpful not only to compare the relative sensitivity to exaggeration of different effort tests but also to study discrepancies between scores on different effort tests applied to the same people. A group of patients which passes TOMM but fails WMT, or another equally easy effort test, is exhibiting gross inconsistency of performance, which we cannot explain on the basis of ability, and it needs to be studied just as closely as the group which fails both of these effort tests.

One approach which is strongly discouraged is the use of clinical judgement alone, with no specific method for evaluating the effort applied to testing [1, 2]. The latter authors [2] point to further sources of divergent conclusions, commenting that some people rely on freely available tests, such as the Rey 15 item test, even though they are quite insensitive in detecting exaggeration. Also, some psychologists conclude that results are invalid only if a person scores significantly worse than the chance level, whereas others use cut-offs based on groups with known severe impairment, such as patients with severe brain injuries. Iverson and Binder [2] argue that the worse than chance criterion, far from being conservative, leads to an unacceptably high failure rate in identifying true cases of exaggeration because the vast majority of people known to be deliberately faking impairment do not score at worse than chance levels. Hence, the hit rate in identifying exaggerators is literally worse than chance, if a worse-than-chance criterion is the only method used.

Let us assume that two psychologists see the same patient at different times, that they use the same symptom validity tests and that the patient fails these tests in both assessments. Even here, there is room for disagreement. Any test will have some false positive classifications of incomplete effort and one psychologist but not the other might conclude that the case is a false positive. In a very few cases, for example, the WMT, the CARB and the TOMM have been failed by people who almost certainly had such severe cognitive impairment that they simply could not score above the cut-offs, despite their best effort. One such case was an 84 year old woman with advanced multi-infarct dementia, whose memory was so poor that she routinely forgot events occurring only a few moments earlier. She scored 65% correct on WMT immediate recognition, 77.3% correct on CARB and 42 correct on TOMM trial two, therefore scoring below the cut-offs on all three effort tests. Her Trail Making A score was 160 seconds. Clinically, such cases can be identified as being very severely impaired based on the diagnosis, on the presence of significant abnormalities on CT or MRI of the brain, on the fact that relatives will not allow them to drive and on obvious dependence on others in day to day life for mobility, self-care and feeding. In these cases, the interpretation of symptom validity test results can be modified accordingly. Such post hoc interpretation does, however, introduce another source of potential disagreement. Where do we draw the line

between someone who is so obviously impaired that they cannot perform almost any test in a reliable way and a person who could pass an effort test but chooses not to do so?

Some might argue that a person failed effort tests because of anxiety or depression during testing. Would this be a reasonable conclusion? Probably not, unless there is strong and independent empirical research to support it and presently such research does not exist. Some studies have reported cognitive impairment in depressed patients (e.g. [19]) but this study, like most others of its type, did not rule out incomplete effort as one source of the apparent impairment on neuropsychological tests. It is important to note that Veiel took a descriptive approach and left open the possibility that poor motivation could have played a significant role in the low test scores obtained from depressed patients in his meta-analytical study. Incomplete effort would be one way of making sense out of the fact that all groups of depressed patients in that study scored a mean of at least 90 seconds on Trail Making B. In contrast, the mean Trail Making B score was only 101 seconds for brain injury patients unable to respond to commands for up to two weeks after head injury [20]. Trail Making B scores in patients with depression comparable to those found in people with brain injuries prompt us to ask whether the depressed patients actually had greater cognitive impairment than people with severe brain injuries or whether incomplete effort in a subgroup inflated their apparent impairment?

One study tested whether, after controlling for inadequate effort, depressive symptoms could explain impaired neuropsychological test results [21]. No such effect was found in 658 consecutive compensation cases undergoing neuropsychological assessment, after those who failed either the CARB or the WMT were excluded from the study. In the 452 cases, who were presumed to be making a valid effort, numerous test results were compared between those with the highest Beck Depression Inventory scores and those with the lowest scores. There were no significant differences between these groups in their results on any of the tests examined, including IQ tests, manual tests, problem solving tests and tests of learning and memory, attention, fluency, working memory, sense of smell and judgment of emotion in tone of voice. Contrary to expectation, reported depressive symptoms had no effect on any of these tests, which do measure ability, and most of the patients in the most depressed group had a diagnosis of major depression. Therefore, it is very unlikely that depression would cause a person to be unable to perform tests like the CARB and WMT, on which the cutoffs for incomplete effort are set at approximately three standard deviations below the mean scores from patients with severe brain injuries [13] and neurological diseases [22]. This does not rule out the possibility that severe depression could lead to impaired judgment and fear of loss of benefits, such that a person might feel that they need to exaggerate in order to prove that they are disabled and in need of disability payments. Nor does it rule out lack of effort because of preoccupation with suicidal concerns or a generalized impairment of initiative and motivation. Nor does it mean that they are not truly disabled by their illness. However, in such cases, the test results would still be invalid as indicators of brain function. Failure on effort tests inevitably implies invalid test results, irrespective of depression.

A further source of disagreement is the clinician's beliefs about how much chronic pain at the time of testing could have affected test performance. Hart, Martelli and Zasler [23], for example, reviewed numerous studies in which patients with chronic pain showed impaired scores on various ability tests. However, almost all of the studies reviewed did not assess or control for the effect of exaggeration of cognitive impairment in suppressing test scores. In the study they quoted, which did measure effort [5], a high rate of exaggeration was found in chronic pain patients and more than 50% of these patients were thought to be producing invalid test results. This led the reviewers [23] to state that further attention to the variable of effort is needed with this population and that effort testing is warranted in clinical assessments and in group studies of patients with chronic pain. Other writers [24] have recommended the use of tests of effort in studies of patients with chronic fatigue syndrome.

The CARB and the WMT were given to two groups of patients diagnosed with fibromyalgia in a pure research study, in which patients were told that their memory test results and other findings would not be entered into their clinical files [6]. One group consisted of people who were still working and/or were not claiming disability, whereas the other was composed of people either already receiving compensation for disability or claiming such compensation. In the no-disability fibromyalgia group, only two cases (4%) failed either the CARB or the WMT. In contrast, 35% of those involved in compensation or disability claims failed one or both of these effort tests. Furthermore, those claiming disability failed more often than those who had already been granted medical disability benefits. The disability and no disability groups had the same diagnosis and

very similar symptoms. There are probably many clinicians who would be tempted to conclude that a patient with fibromyalgia failed an effort test because of pain or emotional distress. The latter study suggests that the critical variable predicting failure on symptom validity tests is disability status, rather than the symptoms of fibromyalgia, such as pain or fatigue.

In principle, many other factors could be argued to impair a person's ability to perform effort tests (e.g. test anxiety, drug usage, confusion, frontal lobe dysfunction, non-fluent English) and this gives rise to differing interpretations in clinical cases. Only by examining each hypothesis with empirical studies will it be possible eventually to resolve conflicts of interpretation and give rise to a more uniform understanding of the effects of psychiatric states and other diagnoses on tests of effort. In the meantime, it needs to be remembered that effort tests, such as the CARB, WMT, TOMM, Portland Digit Recognition Test [3], Amsterdam Short Term Memory Test [5], the Victoria Symptom Validity Test [25] and similar tests are labelled as "effort tests" or "symptom validity tests" because they are easily passed by almost anyone who tries to do so. They have been shown to be very easy for people with severe brain injuries and neurological conditions, such as stroke, multiple sclerosis, brain tumour and ruptured cerebral aneurysms.

In some cases, the cerebral dysfunction will be so great that even simple effort tests exceed the patients' capacity and these cases would represent false positive identifications of incomplete effort, if interpreted out of the clinical context. Nevertheless, in the majority of cases, for whom there is no compelling evidence of widespread and severe cognitive impairment, arising from a known neurological condition, such as a brain tumour or dementia, great caution should be exercised in concluding that a patient performed at a substandard level on effort tests despite making a full effort to do well. In the first stages of research validating tests of effort, a great deal of painstaking work has been done to minimize the occurrence of false positives in samples of patients known to have significant cognitive impairment (i.e. concluding response bias or exaggeration in a person with so much impairment that they actually could not pass the test). Equally rigorous empirical steps need to be taken to reduce false negatives and we should not conclude, based on an untested assumption or a clinical hunch, that a person was unable to pass an effort test because of depression, chronic pain, anxiety or for some other reason, despite trying their best, when they actually chose to fail and were exaggerating. Unless there is empirical evidence from controlled group studies to show that a person's mood state, pain, fatigue or other factors genuinely impair performance on effort tests, it should be concluded that failure on effort testing indicates insufficient effort to produce valid test results. In such cases, symptom exaggeration is invariably present. We might conclude "The patient appeared to be putting forth a good effort and also claimed to be making a full effort but the results are of doubtful validity, owing to the presence of response bias".

Response bias, indicated by scores on a composite symptom validity measure derived from the WMT, CARB and a logit function applied to the California Verbal Learning Test, was found to explain more than 50% of the variance in 43 neuropsychological test results in a sample of 904 compensation claimants [25]. In contrast, years of education, which is known to have an important influence on test scores, explained only 11% of the variance in the same data.. Also, the suppression of test scores produced by severe brain injuries and neurological diseases, such as brain tumors, was very minor in comparison with the effects of cognitive exaggeration. We are learning that, when compensation is potentially available for disability in groups of patients, effort has such a powerful effect on test scores that, in future, it will probably become standard practice to measure and control for effort in group studies, just as in the past, it has been routine to control for important but, in comparison with effort, relatively minor variables, such as age, intelligence and years of education.

Acknowledgement

Dr. Paul R. Lees-Haley and Dr. David Hartman for draft review and critique.

References

[1] J.J. Sweet, Malingering: differential diagnosis, in: *Forensic Neuropsychology Fundamentals and Practice*, J.J. Sweet, Swets and Zeitlinger, Lisse, 1999.

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

[3] L.M. Binder, Assessment of malingering after mild head trauma with the Portland Digit Recognition Test, *Journal of Clinical and Experimental Neuropsychology* **15** (1993), 170–183.

[4] S.P. Van der Werf, J.B. Prins, P.J. Jongen, J.W. van der Meer and G. Bleijenberg, Abnormal neuropsychological findings are not necessarily a sign of cerebral impairment: A matched comparison between chronic fatigue syndrome and multiple sclerosis, *Neuropsychiatry, Neuropsychology and Behavioural Neurology* **13**(3) (2000), 199–203.

[5] B. Schmand, J. Lindeboom, S. Schagen, R. Heijt, T. Koene and H.L. Hamburger, Cognitive complaints in patients after whiplash injury: the impact of malingering, *J Neurol Neurosurg Psychiatry* **64** (1998), 339–343.

[6] R. Gervais, P. Green, A.S. Russell, S. Pieschl and L.M. Allen, Failure on symptom validity tests associated with disability incentives in fibromyalgia patients, *Archives of Clinical Neuropsychology* **15**(8) (2000), 841–842.

[7] L.M. Allen, R.L. Conder, P. Green and D.R. Cox, *CARB' 97 Manual for the Computerized Assessment of Response Bias*, CogniSyst, Inc., Durham, NC, 1997.

[8] L. Flaro, P. Green and L.M. Allen, Symptom validity test results with children: CARB and the WMT, *Archives of Clinical Neuropsychology* **15**(8) (2000), 840.

[9] R. Conder, L. Allen and D. Cox, *Manual for the Computerized Assessment of Response Bias*, CogniSyst, Inc., Durham, NC, 1992.

[10] P. Green, L.M. Allen and K. Astner, The Word Memory Test (Manual): A User's Guide to the Oral and Computer-Administered Forms, US Version 1.1. CogniSyst, Inc., Durham, NC, 1996.

[11] P. Green, G.L. Iverson and L. Allen, Detecting malingering in head injury litigation with the Word Memory Test, *Brain Injury* **13**(10) (1999), 813–819.

[12] P. Green and G. Iverson, Validation of the computerized assessment of response bias in litigating patients with head injuries, *The Clinical Neuropsychologist* (in press).

[13] L.M. Allen and P. Green, Severe traumatic brain injury performance on CARB and the WMT, *Supplement to the CARB and WMT manuals*, Cognisyst, Durham, NC, 1999.

[14] T.N. Tombaugh, *Test of Memory Malingering*, Multi-Health Systems, Toronto, Ontario, 1996.

[15] R. Gervais, P. Green and L.M. Allen, Differential sensitivity to symptom exaggeration of verbal, visual-spatial and numerical symptom validity tests, *Archives of Clinical Neuropsychology* **14**(8) (1999), 746–747.

[16] P. Green, L.M. Allen, J. Berendt and A. Mandel, Relative sensitivity of the Word Memory Test and the Test of Memory Malingering in 144 disability claimants, *Archives of Clinical Neuropsychology* **15**(8) (2000), 841.

[17] D. Williamson, P. Green, Allen and M. Rohling, Who's exaggerating? The Category Test and the Word Memory Test give different answers, *Archives of Clinical Neuropsychology* **15**(8) (2000), 844.

[18] P. Green, L. Allen and R. Gervais, Patterns of memory complaints in two consecutive series of compensation claimants passing or failing symptom validity tests, *Archives of Clinical Neuropsychology* **15**(8) (2000), 844.

[19] H.O.F. Veiel, A preliminary profile of neuropsychological deficits associated with major depression, *Journal of Clinical and Experimental Neuropsychology* **19**(4) (1997), 587–606.

[20] S.S. Dikmen, J.E. Machamer, H.R. Winn and N.R. Temkin, Neuropsychological outcome at 1 year post head injury, *Neuropsychology* **9** (1995), 80–90.

[21] M.L. Rohling, P. Green, L.M. Allen and G. Iverson, Depressive symptoms and neurocognitive test scores in patients passing symptom validity tests, *Archives of Clinical Neuropsychology* (in press).

[22] P. Green and L.M. Allen, Performance of neurological patients on CARB and the WMT, *Supplement to the CARB and WMT manuals*, Cognisyst, Durham, NC, 1999.

[23] R.P. Hart, M.F. Martelli and N.D. Zasler, Chronic pain and neuropsychological functioning, *Neuropsychology Review* **10**(3) (2000), 131–149.

[24] D. Slick, G. Hopp, E. Strauss and G. Thompson, *The Victoria Symptom Validity Test*, PAR, Odessa, Florida, 1997.

[25] M. Rohling, Effect Sizes of Impairment Associated with Symptom Exaggeration versus Definite Traumatic Brain Injury, *Archives of Clinical Neuropsychology* **15**(8) (2000), 843.