Journal of the International Neuropsychological Society (2012), **18**, 625–631.

doi:10.1017/S1355617712000240

DIALOGUE

Performance Validity and Symptom Validity in Neuropsychological Assessment

Glenn J. Larrabee, Ph.D.
Independent Practice, Sarasota, Florida.

(Received November 14, 2011; Final Revision February 7, 2012; Accepted February 8, 2012)

Abstract

Failure to evaluate the validity of an examinee's neuropsychological test performance can alter prediction of external criteria in research investigations, and in the individual case, result in inaccurate conclusions about the degree of impairment resulting from neurological disease or injury. The terms *performance validity* referring to validity of test performance (PVT), and *symptom validity* referring to validity of symptom report (SVT), are suggested to replace less descriptive terms such as effort or response bias. Research is reviewed demonstrating strong diagnostic discrimination for PVTs and SVTs, with a particular emphasis on minimizing false positive errors, facilitated by identifying performance patterns or levels of performance that are atypical for bona fide neurologic disorder. It is further shown that false positive errors decrease, with a corresponding increase in the positive probability of malingering, when multiple independent indicators are required for diagnosis. The rigor of PVT and SVT research design is related to a high degree of reproducibility of results, and large effect sizes of $d = 1.0$ or greater, exceeding effect sizes reported for several psychological and medical diagnostic procedures. (*JINS*, 2012, *18*, 625–631)

Keywords: Malingering, False positive errors, Sensitivity, Specificity, Diagnostic probability, Research design

Bigler acknowledges the importance of evaluating the validity of examinee performance, but raises concerns about the meaning of "effort," the issue of what "near pass" performance means (i.e., scores that fall just within the range of invalid performance), the possibility that such scores may represent "false positives," and the fact that there are no systematic lesion localization studies of Symptom Validity Test (SVT) performance. Bigler also discusses the possibility that illness behavior and "diagnosis threat" (i.e., the influence of expectations) can affect performance on SVTs. He further questions whether performance on SVTs may be related to actual cognitive abilities and to the neurobiology of drive, motivation and attention. Last, he raises concerns about the rigor of existing research underlying the development of SVTs.

Bigler and I are in agreement about the need to assess the validity of an examinee's performance. Failure to do so can lead to misleading results. My colleagues and I (Rohling et al., 2011) reviewed several studies in which performance on response bias indicators (another term for SVTs) attenuated the correlation between neuropsychological test measures and an external criterion. For example, grade point average and Full Scale IQ correlated below the expected level of strength until those subjects failing an SVT were excluded (Greiffenstein & Baker, 2003); olfactory identification was only correlated with measures of brain injury severity (e.g., Glasgow Coma Scale) for those subjects passing an SVT (Green, Rohling, Iverson, & Gervais, 2003); California Verbal Learning Test scores did not discriminate traumatic brain injury patients with abnormal CT or MRI scans from those with normal scans until patients failing SVTs were excluded (Green, 2007); patients with moderate or severe traumatic brain injury (TBI), 88% of whom had abnormal CT or MRI, plus patients with known cerebral impairment (stroke, aneurysm) did not differ from those with uncomplicated mild TBI, psychiatric disorders, orthopedic injuries or chronic pain until those failing an SVT were dropped from comparison (Green, Rohling, Lees-Haley, & Allen, 2001). An association between memory complaints and performance on the California Verbal Learning Test, which goes counter to the oft-replicated finding of no association between memory or cognitive complaints and actual test performance (Brulot, Strauss, & Spellacy, 1997; Hanninen et al., 1994; Larrabee & Levin, 1986; Williams, Little, Scates, & Blockman, 1987), disappeared when those failing an SVT were excluded (Gervais, Ben-Porath,

Correspondence and reprint requests to: Glenn J. Larrabee, 2650 Bahia Vista Street, Suite 308, Sarasota, FL 34239. E-mail: glarrabee@aol.com

Wygant, & Green, 2008). Subsequent to the Rohling et al. (2011) review, Fox (2011) showed that the association between neuropsychological test performance and presence/absence of brain injury only was demonstrated in patients passing SVTs.

Some of the debate regarding symptom validity testing results from use of the term "effort." "Effort" suggests a continuum, ranging from excellent, to very poor. SVTs are constructed, however, based on patterns of performance that are atypical in either pattern or degree, in comparison to the performance of patients with bona fide neurologic disorder. Consequently, SVTs demonstrate a discontinuity in performance rather than a continuum of performance, with most neurologic patients either not showing the atypical pattern, or performing at ceiling on a particular SVT. Examples of atypical patterns of performance include poorer performance on measures of attention than on measures of memory (Mittenberg, Azrin, Millsaps, & Heilbronner, 1993), or poorer performance on gross compared to fine motor tests (Greiffenstein, Baker, & Gola, 1996). Examples of atypical degree include motor function performance at levels rarely seen in patients with severe neurologic dysfunction (Greiffenstein, 2007). Consequently, performance is so atypical for bona fide neurologic disease that persons with significant neurologic disorders rarely fail "effort" tests. For example, the meta-analysis of Vickery, Berry, Inman, Harris, & Orey, 2001, reported a 95.7% specificity or 4.3% false positive rate. Additionally, modern SVT research typically sets specificity at 90% or better on individual tests, yielding a false positive rate of 10% or less (Boone, 2007; Larrabee, 2007; Morgan & Sweet, 2009). Consequently, if SVTs are unlikely to be failed by persons with significant neurologic dysfunction, then performance on these tasks actually requires very minimal levels of neurocognitive capacity and consequently, very little "effort." As a result, I have recently advocated for referring to SVTs as measures of *performance validity* to clarify the extent to which a person's test performance is or is not an accurate reflection of their actual level of ability (Bigler, Kaufmann, & Larrabee, 2010; Larrabee, 2012). This term is much more descriptive than the terms "effort, "symptom validity," or "response bias," and in keeping with the longstanding convention of psychologists commenting on the validity of test results. Moreover, the term "symptom validity" is actually more descriptive of *subjective complaint* than it is of *performance*. Thus, I recommend that we use two descriptive terms in evaluating the validity of an examinee's neuropsychological examination: (1) *performance validity* to refer to the validity of actual ability task performance, assessed either by stand-alone tests such as Dot Counting or by atypical performance on neuropsychological tests such as Finger Tapping, and (2) *symptom validity* to refer to the accuracy of symptomatic complaint on self-report measures such as the MMPI-2.

As previously noted, false positive rates are typically 10% or less on individual Performance Validity Tests (PVTs). For example, the manual for the Test of Memory Malingering (TOMM; Tombaugh, 1996) contains detailed information about the performance of aphasic, TBI, dementia, and neurologic patients, very few of whom (with the exception of dementia) perform below the recommended cutoff. Three patients with severe anoxic encephalopathy and radiologically confirmed hippocampal damage scored in the valid performance range on the recognition trials of the Word Memory Test (Goodrich-Hunsaker & Hopkins, 2009). Similarly, psychiatric disorders have not been found to impact PVT scores, including depression (Rees, Tombaugh, & Boulay, 2001), depression and anxiety (Ashendorf, Constantinou, & McCaffrey, 2004), and depression with chronic pain (Iverson, Le Page, Koehler, Shojania, & Badii, 2007).

Illness behavior and "diagnosis threat" do not appear to impact PVT scores. Acute pain (cold pressor) has no impact on performance on Reliable Digit Span (Etherton, Bianchini, Ciota, & Greve, 2005) or on the TOMM (Etherton, Bianchini, Greve, & Ciota, 2005). Suhr and Gunstad (2005) did not find differences on the WMT for those mild TBI subjects in the "diagnosis threat" condition compared to those in the non-threat group. Arguments that neurological mechanisms related to drive and motivation underlie PVT performance are not supported in light of PVT profiles which are typically valid for patients with significant neurologic disease due to diverse causes, showing that these tasks require little in the way of effort, drive or motivation and, as mentioned, general neurocognitive capacity; that is, performance is usually near ceiling even in contexts of severe objectively verified cerebral disorders. For example, on TOMM Trial 2, 21 aphasic patients averaged 98.7% correct, and 22 TBI patients (range of 1 day to 3 months of coma) averaged 98.2% correct; indeed, one patient with gunshot wound, right frontal lobectomy, and 38 days of coma scored 100% on Trial 2 (Tombaugh, 1996). In this vein, a patient with significant abulia due to severe brain trauma, who would almost certainly require 24-hr supervision, and be minimally testable from a neuropsychological standpoint, would not warrant SVT or PVT assessment. In such a patient, there would of course be a legitimate concern about false positive errors on PVTs. As with any mental test, consideration of context is necessary and expected.

One of Bigler's major concerns, the "near pass" (i.e., performance falls just within the invalid range on a PVT), is not restricted to PVT investigations, it is a pervasive concern in the field of assessment. One's child does or does not reach the cutoff to qualify for the gifted class or for help for a learning disability. One's neuropsychological test score does or does not reach a particular level of normality/abnormality (Heaton, Miller, Taylor, & Grant, 2004). Current PVT research focuses on avoiding the error of misidentifying as invalid the performance of a patient with a bona fide condition who is actually producing a valid performance. Moreover, there is a strong statistical argument for keeping false positive errors at a minimum: Positive Predictive Power (PPP), or the probability of a diagnosis, is more dependent upon Specificity (accurately diagnosing a person without the target disorder as not having the disorder) than Sensitivity (correctly identifying persons with the target disorder as having the disorder; see Straus, Richardson, Glasziou, & Haynes, 2005). Since the basic formula for PPP is (True Positives) ÷ (True Positives + False positives), the PVT investigator attempts to keep false positives at a minimum. As noted in the previous meta-analysis (Vickery, Berry, Inman, Harris, & Orey, 2001) as well as in recent reviews

(Boone, 2007; Larrabee, 2007), false positives are typically 10% or less, with much lower sensitivities (56% per Vickery et al., 2001). Researchers also advocate reporting the characteristics of subjects identified as false positive cases in individual PVT investigations (Larrabee, Greiffenstein, Greve, & Bianchini, 2007; also see Victor, Boone, Serpa, Buehler, & Ziegler, 2009). This clarifies the characteristics of those patients with truly severe mental disorders who fail PVTs on the basis of actual impairment. This information provides the clinician with concrete injury/clinical fact patterns that legitimately correlate with PVT failure, thereby facilitating individual comparisons on a case by case basis (e.g., coma and structural lesions in the brain; Larrabee, 2003a; unequivocally severe and obvious neurologic symptoms, Merten, Bossink, & Schmand, 2007; or need for 24-hr supervision; Meyers & Volbrecht, 2003). Authors of PVTs have also included comparison groups with various neurologic, psychiatric and developmental conditions to further reduce the chances of false positive identification on an individual PVT (Boone, Lu, & Herzberg, 2002a, 2002b; Tombaugh, 1996).

PVTs have two applications in current neuropsychological practice: (1) screening data for a research investigation to remove effects of invalid performance (see Larrabee, Millis, & Meyers, 2008) and (2) for evaluation of an individual patient to determine if performance of that patient is valid, and forensically, to address the issue of malingering. Concerns about false positives are of far greater import in the second application of PVTs, since there really is no consequence to the patient whose data are excluded from clinical research.

Concerns about false positive identification ("near passes") in the individual case are addressed by the diagnostic criteria for Malingered Neurocognitive Dysfunction (MND; Slick, Sherman, & Iverson, 1999). Slick et al. define malingering as the volitional exaggeration or fabrication of cognitive dysfunction for the purpose of obtaining substantial material gain, avoiding or escaping legally mandated formal duty or responsibility. Criteria for MND require a substantial external incentive (e.g., litigation, criminal prosecution), multiple sources of evidence from behavior (e.g., PVTs), and symptom report (e.g., SVTs) to define probable malingering, whereas significantly worse-than-chance performance defines definite malingering. Moreover, these sources of evidence must not be the product of neurological, psychiatric or developmental factors (note the direct relevance of this last criterion to the issue of false positives).

The Slick et al. criteria for MND have led to extensive subsequent research using these criteria for known-group investigations of detection of malingering (Boone, 2007; Larrabee, 2007; Morgan and Sweet, 2009). These criteria have also influenced development of criteria for Malingered Pain Related Disability (MPRD; Bianchini, Greve, and Glynn (2005). As my colleagues and I have pointed out (Larrabee et al., 2007), the diagnostic criteria for MND and MPRD share key features: (1) the requirement for a substantial external incentive, (2) the requirement for multiple indicators of performance invalidity or symptom exaggeration, and (3) test performance and symptom report patterns that are atypical in pattern and degree for bona fide neurologic, psychiatric or developmental disorders. It is the combined improbability of findings, in the context of external incentive, without any viable alternative explanation, that establishes the intent of the examinee to malinger (Larrabee et al., 2007).

Research using the Slick et al. MND criteria shows the value of requiring multiple failures on PVTs and SVTs to determine probabilities of malingering in contexts with substantial external incentives. I (Larrabee, 2003a) demonstrated that requiring failure of two embedded/derived PVTs and/or SVTs resulted in a sensitivity of .875 and specificity of .889 for discriminating litigants (primarily with uncomplicated MTBI) performing significantly worse than chance from clinical patients with moderate and severe TBI. The requirement that patients fail 3 or more PVTs and SVTs resulted in a sensitivity of .542, but a specificity of 1.00 (i.e., there were no false positives). These data were replicated by Victor et al. (2009) using a different set of embedded/derived indictors in a similar research design yielding sensitivity of .838 and specificity of .939 for failure of any two PVTs, and sensitivity of .514 and specificity of .985 for failure of three or more PVTs.

The drop in false alarm rate and increase in specificity going from two to three failed PVTs/SVTs, is directly related to the PPP of malingering, as demonstrated by the methodology of *chaining* likelihood ratios (Grimes & Schulz, 2005). The positive likelihood ratio is defined by the ratio of sensitivity to the false positive rate. Hence, a score falling at a particular PVT cutoff that has an associated sensitivity of .50 and specificity of .90 would yield a likelihood ratio of .50 ÷ .10, or 5.0. If this is then premultiplied by the base rate odds of malingering (assume a malingering base rate of .40, per Larrabee, Millis, & Meyers, 2009, yielding a base rate odds of (base rate) ÷ (1 – base rate) or (.40) ÷ (1–.40) or .67), this value becomes .67 × 5.0 or 3.35. This can be converted back to a probability of malingering by the formula (odds) ÷ (odds + 1), in this case, 3.35 ÷ 4.35, or .77. If the indicators are independent, they can be chained, so that the post-test odds after premultiplying one indicator by the base rate odds, become the new pretest odds by which a second independent indicator is multiplied. Thus, if a second PVT is failed at the same cut off yielding sensitivity of .50 and specificity of .90, this yields a second likelihood ratio of 5.0, which is now multiplied by the post-test odds of 3.35 obtained after failure of the first indicator. This yields new post-test odds of 16.75, which can be converted back to a probability by dividing 16.75 by 17.75 to yield a probability of malingering of .94, in settings with substantial external incentive. The interested reader is referred to a detailed explanation of this methodology (Larrabee, 2008).

The method of chaining of likelihood ratios shows how the probability of confidently determining malingered performance is enhanced by requiring multiple PVT and SVT failure, consistent with other results (Larrabee, 2003a; Victor et al., 2009). Boone and Lu (2003) make a related point regarding the decline in false positive rate by using several independent tests, each with a false positive rate of .10: failure of two PVTs yields a probability (false positive rate) of .01 (.1 × .1), whereas failure of

https://doi.org/10.1017/S1355617712000240 Published online by Cambridge University Press

three PVTs yields a probability of .001 (.1 × .1 × .1), and failure of six PVTs yields a probability as low as one in a million (.1 × .1 × .1 × .1 × .1 × .1). Said differently, the standard of multiple PVT and SVT failure protects against false positive diagnostic errors. Per Boone and Lu (2003), Larrabee (2003a; 2008), and Victor et al. (2009), failure of three independent PVTs is associated with essentially no false positive errors, a highly compelling empirically-based conclusion in the context of *any* form of diagnostic testing.

PVT performance can vary in persons identified with multiple PVT failures and should be assessed throughout an examination (Boone, 2009). Malingering can lower performance as much as 1 to 2 *SD* on select sensitive tests of memory and processing speed (Larrabee, 2003a), and PVT failure can lower the overall test battery mean (Green et al., 2001) by over 1 *SD*. In the presence of malingering, poor performances are more likely the result of intentional underperformance, particularly in conditions such as uncomplicated mild TBI in which pronounced abnormalities are unexpected (McCrea et al., 2009), and normal range performances themselves are likely underestimates of actual level of ability.

Last, there is a lengthy history of strong experimental design in PVT and SVT investigations. Research designs in malingering are discussed over 20 years ago in Rogers' first book (Rogers, 1988). The two most rigorous and clinically relevant designs are the simulation design, and the "known groups" or "criterion group" designs (Heilbronner, et al., 2009). In the simulation design, a non-injured group of subjects is specifically instructed to feign deficits on PVTs, SVTs, and neuropsychological ability tests, which are then contrasted with scores produced by a group of persons with bona fide disorders, usually patients with moderate or severe TBI. The resulting patterns discriminate known feigning from legitimate performance profiles associated with moderate and severe TBI, thereby minimizing false positive diagnosis in the TBI group. The disadvantage is that issues arise as to the "real world" generalization of non-injured persons feigning deficit compares to actual malingerers who have significant external incentives, for example, millions at stake in a personal injury claim. The second design, criterion groups, contrasts the performance of a group of litigating subjects, usually those with alleged non-complicated mild TBI, who have failed multiple PVTs and SVTs, commonly using the Slick et al. MND criteria, with a group of clinical patients, typically with moderate and severe TBI. This has the advantage of using a group with "real world" incentives, that is unlikely to have significant neurological damage and persistent neuropsychological deficits (McCrea et al., 2009), holding false positive errors at a minimum by determining performance patterns that are not characteristic of moderate and severe TBI. Although random assignment cannot be used for the simulation and criterion group designs just described, these designs are appropriate for case control comparisons.

PVT and SVT research using simulation and criterion group designs has, for the most part, yielded very consistent and replicable findings. For example, Heaton, Smith, Lehman, and Vogt (1978) reported an average dominant plus non-dominant Finger Tapping score of 63.1 for a sample of simulators, which was essentially identical to the score of 63.0 for the simulators in Mittenberg, Rotholc, Russell, and Heilbronner (1996). In a criterion groups design, I reported an optimal dominant plus non-dominant hand Finger Tapping score of less than 63 for discriminating subjects with definite MND from patients with moderate or severe TBI (Larrabee, 2003a), which was identical to the cutting score one would obtain by combining the male and female subjects in the criterion groups investigation of Arnold et al. (2005). In a criterion groups design, I (Larrabee, 2003b) reported optimal MMPI-2 FBS Symptom Validity cutoffs of 21 or 22 in discriminating subjects with definite MND from those with moderate or severe TBI, which was identical to the value of 21 or 22 for discriminating subjects with probable MND from patients with moderate or severe TBI reported by Ross, Millis, Krukowski, Putnam, and Adams (2004). As already noted, Victor et al. (2009) obtained very similar sensitivities and specificities for failure of any two or any three or more PVTs to the values I obtained for failure of any two or three or more PVTs or SVTs (Larrabee, 2003a).

My colleagues and I have relied upon the similarity of findings in simulation and criterion group designs to link together research supporting the psychometric basis of MND criteria (Larrabee et al., 2007). The similarity of findings on individual PVTs for simulators and for litigants with definite MND (defined by worse than chance performance) demonstrates that worse-than-chance performance reflects intentional underperformance; in other words, the definite MND subjects performed identically to non-injured persons feigning impairment who are known to be intentionally underperforming because they have been instructed to do so. Additionally, the PVT and neuropsychological test performance of persons with probable MND (defined by multiple PVT failure independent of the particular PVT or neuropsychological test data being compared) did not differ from that of persons with definite MND, establishing the validity of the probable MND criteria. Last, the paper by Bianchini, Curtis, and Greve (2006) showing a dose-effect relationship between PVT failure and amount of external incentive, supports that intent is causally related to PVT failure.

In closing, the science behind measures of performance and symptom validity is rigorous, well developed, replicable and specifically focused on keeping false positive errors at a minimum. I have also argued for a change in terminology that may reduce some of the confusion in this area, recommending the use of Performance Validity Test (PVT) for measures directed at assessing the validity of a person's performance, and Symptom Validity Test (SVT) for measures directed at assessing the validity of a person's symptomatic complaint.

ACKNOWLEDGMENT

This manuscript has not been previously published electronically or in print. Portions of this study were presented as a debate at the 38th Annual International Neuropsychological Society meeting on "Admissibility and Appropriate Use of Symptom Validity Science in Forensic Consulting" in Acapulco, Mexico, February, 2010,

moderated by Paul M. Kaufmann, J.D., Ph.D. Dr. Larrabee is engaged in the full time independent practice of clinical neuropsychology with a primary emphasis in forensic neuropsychology. He is the editor of *Assessment of Malingered Neuropsychological Deficits* (2007, Oxford University Press), and *Forensic Neuropsychology. A Scientific Approach* (2nd Edition, 2012, Oxford University Press), and receives royalties from the sale of these books.

REFERENCES

Arnold, G., Boone, K.B., Lu, P., Dean, A., Wen, J., Nitch, S., & McPherson, S. (2005). Sensitivity and specificity of Finger Tapping Test scores for the detection of suspect effort. *The Clinical Neuropsychologist*, *19*, 105–120.

Ashendorf, L., Constantinou, M., & McCaffrey, R.J. (2004). The effect of depression and anxiety on the TOMM in community-dwelling older adults. *Archives of Clinical Neuropsychology*, *19*, 125–130.

Bianchini, K.J., Curtis, K.L., & Greve, K.W. (2006). Compensation and malingering in traumatic brain injury: A dose-response relationship? *The Clinical Neuropsychologist*, *20*, 831–847.

Bianchini, K.J., Greve, K.W., & Glynn, G. (2005). On the diagnosis of malingered pain-related disability: Lessons from cognitive malingering research. *Spine Journal*, *5*, 404–417.

Bigler, E.D., Kaufmann, P.M., & Larrabee, G.J. (2010, February). Admissibility and appropriate use of symptom validity science in forensic consulting. Talk presented at the 38th Annual Meeting of the International Neuropsychological Society, Acapulco, Mexico.

Boone, K.B. (2007). *Assessment of feigned cognitive impairment. A neuropsychological perspective*. New York: Guilford.

Boone, K.B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist*, *23*, 729–741.

Boone, K.B., & Lu, P.H. (2003). Noncredible cognitive performance in the context of severe brain injury. *The Clinical Neuropsychologist*, *17*, 244–254.

Boone, K.B., Lu, P., & Herzberg, D.S. (2002a). *The b Test. Manual*. Los Angeles, CA: Western Psychological Services.

Boone, K.B., Lu, P., & Herzberg, D.S. (2002b). *The Dot Counting Test. Manual*. Los Angeles, CA: Western Psychological Services.

Brulot, M.M., Strauss, E., & Spellacy, F. (1997). Validity of the Minnesota Multiphasic Personality Inventory-2 correction factors for use with patients with suspected head injury. *The Clinical Neuropsychologist*, *11*, 391–401.

Etherton, J.L., Bianchini, K.J., Ciota, M.A., & Greve, K.W. (2005). Reliable Digit Span is unaffected by laboratory-induced pain: Implications for clinical use. *Assessment*, *12*, 101–106.

Etherton, J.L., Bianchini, K.J., Greve, K.W., & Ciota, M.A. (2005). Test of memory malingering performance is unaffected by laboratory-induced pain: Implications for clinical use. *Archives of Clinical Neuropsychology*, *20*, 375–384.

Fox, D.D. (2011). Symptom validity test failure indicates invalidity of neuropsychological tests. *The Clinical Neuropsychologist*, *25*, 488–495.

Gervais, R.O., Ben-Porath, Y.S., Wygant, D.B., & Green, P. (2008). Differential sensitivity of the Response Bias Scale (RBS) and MMPI-2 validity scales to memory complaints. *The Clinical Neuropsychologist*, *22*, 1061–1079.

Goodrich-Hunsaker, N.J., & Hopkins, R.O. (2009). Word Memory Test performance in amnesic patients with hippocampal damage. *Neuropsychology*, *23*, 529–534.

Green, P. (2007). The pervasive influence of effort on neuropsychological tests. *Physical Medicine and Rehabilitation Clinics of North America*, *18*, 43–68.

Green, P., Rohling, M.L., Iverson, G.L., & Gervais, R.O. (2003). Relationships between olfactory discrimination and head injury severity. *Brain Injury*, *17*, 479–496.

Green, P., Rohling, M.L., Lees-Haley, P.R., & Allen, L.M. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury*, *15*, 1045–1060.

Greiffenstein, M.F. (2007). Motor, sensory, and perceptual-motor pseudoabnormalities. In G.J. Larrabee (Ed.). *Assessment of malingered neuropsychological deficits* (pp. 100–130). New York: Oxford University Press.

Greiffenstein, M.F., & Baker, W.J. (2003). Premorbid Clues? Preinjury scholastic performance and present neuropsychological functioning in late postconcussion syndrome. *The Clinical Neuropsychologist*, *17*, 561–573.

Greiffenstein, M.F., Baker, W.J., & Gola, T. (1996). Motor dysfunction profiles in traumatic brain injury and post-concussion syndrome. *Journal of the International Neuropsychological Society*, *2*, 477–485.

Grimes, D.A., & Schulz, K.F. (2005). Epidemiology 3. Refining clinical diagnosis with likelihood ratios. *Lancet*, *365*, 1500–1505.

Hanninen, T., Reinikainen, K.J., Helkala, E.-L., Koivisto, K., Mykkanen, L., Laakso, M., … Riekkinen, R.J. (1994). Subjective memory complaints and personality traits in normal elderly subjects. *Journal of the American Geriatric Society*, *42*, 1–4.

Heaton, R.K., Miller, S.W., Taylor, M.J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults. Professional Manual*. Lutz, FL: Psychological Assessment Resources.

Heaton, R.K., Smith, H.H., Jr., Lehman, R.A., & Vogt, A.J. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, *46*, 892–900.

Heilbronner, R.L., Sweet, J.J., Morgan, J.E., Larrabee, G.J., Millis, S.R., & Conference Participants (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, *23*, 1093–1129.

Iverson, G.L., Le Page, J., Koehler, B.E., Shojania, K., & Badii, M. (2007). Test of Memory Malingering (TOMM) scores are not affected by chronic pain or depression in patients with fibromyalgia. *The Clinical Neuropsychologist*, *21*, 532–546.

Larrabee, G.J. (2003a). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist*, *17*, 410–425.

Larrabee, G.J. (2003b). Detection of symptom exaggeration with the MMPI-2 in litigants with malingered neurocognitive dysfunction. *The Clinical Neuropsychologist*, *17*, 54–68.

Larrabee, G.J. (Ed.) (2007). *Assessment of malingered neuropsychological deficits*. New York: Oxford.

Larrabee, G.J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist*, *22*, 666–679.

Larrabee, G.J. (2012). Assessment of malingering. In G.J. Larrabee (Ed.). *Forensic neuropsychology: A scientific approach* (2nd ed., pp. 116–159). New York: Oxford University Press.

Larrabee, G.J., Greiffenstein, M.F., Greve, K.W., & Bianchini, K.J. (2007). Refining diagnostic criteria for malingering. In G.J. Larrabee (Ed.). *Assessment of malingered neuropsychological deficits* (pp. 334–371). New York: Oxford.

Larrabee, G.J., & Levin, H.S. (1986). Memory self-ratings and objective test performance in a normal elderly sample. *Journal of Clinical and Experimental Neuropsychology*, *8*, 275–284.

Larrabee, G.J., Millis, S.R., & Meyers, J.E. (2008). Sensitivity to brain dysfunction of the Halstead-Reitan vs. an ability-focused neuropsychological battery. *The Clinical Neuropsychologist*, *22*, 813–825.

Larrabee, G.J., Millis, S.R., & Meyers, J.E. (2009). 40 plus or minus 10, a new magical number: Reply to Russell. *The Clinical Neuropsychologist*, *23*, 746–753.

McCrea, M., Iverson, G.L., McAllister, T.W., Hammeke, T.A., Powell, M.R., Barr, W.B., & Kelly, J.P. (2009). An integrated review of recovery after mild traumatic brain injury (MTBI): Implications for clinical management. *The Clinical Neuropsychologist*, *23*, 1368–1390.

Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology*, *29*, 308–318.

Meyers, J.E., & Volbrecht, M.E. (2003). A validation of multiple malingering detection methods in a large clinical sample. *Archives of Clinical Neuropsychology*, *18*, 261–276.

Mittenberg, W., Azrin, R., Millsaps, C., & Heilbronner, R. (1993). Identification of malingered head injury on the Wechsler Memory Scale-Revised. *Psychological Assessment*, *5*, 34–40.

Mittenberg, W., Rotholc, A., Russell, E., & Heilbronner, R. (1996). Identification of malingered head injury on the Halstead-Reitan Battery. *Archives of Clinical Neuropsychology*, *11*, 271–281.

Morgan, J.E., & Sweet, J.J. (Eds.), (2009) *Neuropsychology of malingering casebook*. New York: Psychology Press.

Rees, L.M., Tombaugh, T.N., & Boulay, L. (2001). Depression and the Test of Memory Malingering. *Archives of Clinical Neuropsychology*, *16*, 501–506.

Rogers, R. (1988). Researching dissimulation. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 309–327). New York: Guilford Press.

Rohling, M.L., Larrabee, G.J., Greiffenstein, M.F., Ben-Porath, Y.S., Lees-Haley, P., Green, P., & Greve, K.W. (2011). A misleading review of response bias: Comment on McGrath, Mitchell, Kim, and Hough (2010). *Psychological Bulletin*, *137*, 708–712.

Ross, S.R., Millis, S.R., Krukowski, R.A., Putnam, S.H., & Adams, K.M. (2004). Detecting probable malingering on the MMPI-2: An examination of the Fake-Bad Scale in mild head injury. *Journal of Clinical and Experimental Neuropsychology*, *26*, 115–124.

Slick, D.J., Sherman, E.M.S., & Iverson, G.L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, *13*, 545–561.

Straus, S.E., Richardson, W.S., Glasziou, P., & Haynes, R.B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). New York: Elsevier Churchill Livingstone.

Suhr, J.A., & Gunstad, J. (2005). Further exploration of the effect of "diagnosis threat" on cognitive performance in individuals with mild head injury. *Journal of the International Neuropsychological Society*, *11*, 23–29.

Tombaugh, T.N. (1996). *TOMM. Test of Memory Malingering*. New York: Multi-Health Systems.

Vickery, C.D., Berry, D.T.R., Inman, T.H., Harris, M.J., & Orey, S.A. (2001). Detection of inadequate effort on neuropsychological testing: A meta-analytic review of selected procedures. *Archives of Clinical Neuropsychology*, *16*, 45–73.

Victor, T.L., Boone, K.B., Serpa, J.G., Buehler, J., & Ziegler, E.A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist*, *23*, 297–313.

Williams, J.M., Little, M.M., Scates, S., & Blockman, N. (1987). Memory complaints and abilities among depressed older adults. *Journal of Consulting and Clinical Psychology*, *55*, 595–598.

doi:10.1017/S1355617712000392

DIALOGUE RESPONSE

Response to Bigler

Glenn J. Larrabee

Bigler (this issue) and I apparently are in agreement about the importance of symptom validity testing, and my recommendation to adopt a new terminology of "performance validity" to address the validity of performance on measures of ability, and "symptom validity" to address the validity of symptom report on measures such as the MMPI-2. We appear to differ on issues related to false positives and the rigor of performance and symptom validity research designs.

The study by Locke, Smigielski, Powell, and Stevens (2008) is cited by Bigler as demonstrating potential false positive errors due to TOMM scores falling in a "near miss" zone just below cutoff. This interpretation suggests a continuum of performance. Review of Bigler's Figure 1 and Locke et al.'s Table 2 shows that the frequency distribution of TOMM scores does not, however, reflect a continuum but shows two discrete distributions: (1) a sample of 68 ranging from 45 to 50 (mean = 49.31, $SD = 1.16$) and (2) a sample of 19 ranging from 22 to 44 (mean = 35.11, $SD = 6.55$) [note Bigler interprets two distributions below 45, but the sample size is too small to establish this presence]. Clearly,