False-Positive Rates Associated with the Use of Multiple Performance and Symptom Validity Tests

Glenn J. Larrabee*

Independent Practice, Sarasota, FL, USA

*Corresponding author at: 2650 Bahia Vista St., Suite 308, Sarasota, FL 34239, USA. Tel.: +941-955-9596; fax: +941-957-3485.
E-mail address: glarrabee@aol.com (G.J. Larrabee).

Accepted 31 March 2014

Abstract

Performance validity test (PVT) error rates using Monte Carlo simulation reported by Berthelson and colleagues (in False positive diagnosis of malingering due to the use of multiple effort tests. Brain Injury, 27, 909–916, 2013) were compared with PVT and symptom validity test (SVT) failure rates in two nonmalingering clinical samples. At a per-test false-positive rate of 10%, Monte Carlo simulation overestimated error rates for: (i) failure of ≥2 out of 5 PVTs/SVT for Larrabee (in Detection of malingering using atypical performance patterns on standard neuropsychological tests. The Clinical Neuropsychologist, 17, 410–425, 2003) and ACS (Pearson, Advanced clinical solutions for use with WAIS-IV and WMS-IV. San Antonio: Pearson Education, 2009) and (ii) failure of ≥2 out of 7 PVTs/SVT for Larrabee (Detection of malingering using atypical performance patterns on standard neuropsychological tests. The Clinical Neuropsychologist, 17, 410–425, 2003; Malingering scales for the Continuous Recognition Memory Test and Continuous Visual Memory Test. The Clinical Neuropsychologist, 23, 167–180, 2009 combined). Monte Carlo overestimation is likely because PVT performances are atypical in pattern or degree for what occurs in actual neurologic, psychiatric, or developmental disorders. Consequently, PVT scores form skewed distributions with performance at ceiling and restricted range, rather than forming a standard normal distribution with mean of 0 and standard deviation of 1.0. These results support the practice of using ≥2 PVT/SVT failures as representing probable invalid clinical presentation.

Keywords: Malingering/symptom validity testing; Forensic neuropsychology; Assessment

Proposed diagnostic criteria for probable malingering require multiple indicators of invalid test performance and symptom report and the presence of a strong external incentive (Slick, Sherman, & Iverson, 1999). These criteria also require ruling out developmental, neurologic, and psychiatric explanations for failure of performance and symptom validity measures, to reduce the likelihood of false-positive identification for persons with actual valid clinical presentations. Empirical studies have provided statistical support for these proposed criteria (Larrabee, 2003; Victor, Boone, Serpa, Buehler, & Ziegler, 2009).

Maintaining low false-positive identification rates on measures of performance validity (PVTs such as the Test of Memory Malingering) and symptom validity (SVTs such as the validity scales of the MMPI-2-RF, see Larrabee, 2012) is important in two respects: (i) the false-positive error of concluding malingering, based on a PVT or SVT failure, when malingering is not present can have very significant consequences for the examinee and (ii) the positive predictive power (PPP) of a diagnosis of malingering is dependent upon a low false-positive rate. This second point is demonstrated by the following general formula for PPP: True Positives/(True Positives + False Positives) (Baldessarini, Finklestein, & Arana, 1983). Hence, the PPP increases with decreases in the false-positive rate. Moreover, at cutting scores associated with a false-positive rate of zero that are still exceeded by some proportion of a definite or probable malingering sample, PPP = True Positives/(True Positives + zero False Positives), yielding a 1.00 probability of malingering (in the context of a substantial external incentive).

Other formulae that are used to generate PPP include the application of positive likelihood ratios, which are determined by dividing sensitivity by the false-positive rate; such ratios show the likelihood that a particular score comes from a group with the condition of interest (COI; in the example of malingering, a group defined as malingering) versus the chance it comes from a nonmalingering (i.e., clinical patient) group (Grimes & Schulz, 2005). When the positive likelihood ratio for a particular PVT
or SVT is premultiplied by the base rate odds of malingering, the resulting value represents the post-test odds of malingering, which can be converted back to a PPP by the formula: odds/(odds + 1) (see Larrabee, 2008).

Thus, false-positive rates play a central role in the diagnosis of any condition of interest, with a particularly important role in the diagnosis of malingering. Consequently, cutting scores on individual PVTs and SVTs are set so that specificity (the proportion of nonmalingering subjects correctly identified as nonmalingering) is typically 0.90 or higher, keeping the false-positive rate at 10% or less (Boone, 2007; Larrabee, 2007; Vickery, Berry, Inman, Harris, & Orey, 2001). As a result of conservative cut-offs that minimize false-positive identification, there is a corresponding lower sensitivity (the proportion of malingers correctly identified as malingering), with a higher false-negative rate per-PVT or SVT. This is reflected in Vickery and colleagues’ (2001) meta-analysis, which found a mean specificity of 0.957, with a much lower mean sensitivity of 0.56. In an updated meta-analysis, Sollman and Berry (2011) reported a mean specificity of 0.90 with mean sensitivity of 0.69. Thus, scores on individual PVTs and SVTs are better at ruling in malingering than they are at ruling out malingering. Importantly, as Larrabee (2003, 2008) and Victor and colleagues (2009) have shown, requiring multiple PVT and SVT failures improves the detection of feigning, without appreciably altering the per-test false-positive rate.

Recognizing the importance of false-positive errors, and the requirement for multiple PVTs and SVTs in Slick and colleagues’ (1999) criteria, Berthelson, Mulchan, Odland, Miller, and Mittenberg (2013) undertook an investigation of false-positive rates by determining (i) the average intercorrelation of PVTs based on published research and (ii), using this information, in combination with Monte Carlo simulation, to determine the false-positive rates associated with various numbers of tests failed, as a function of the number of tests administered, and utilizing two per-test false-positive error rates: 10% and 15%. Monte Carlo simulation utilizes randomly generated data for a predetermined number of variables, at a specified level of variable intercorrelation, which yields mean values of 0 and standard deviations of 1.0 for each simulated variable (Crawford, Garthwaite, & Gault, 2007). One can then set a predetermined cut-off for abnormality, for example, −1.282 for the bottom 10%, and determine the percentage of simulated cases that exceed this cut-off for various subsets of the total number of variables. Employing Monte Carlo simulation, Berthelson and colleagues (2013) concluded that the false-positive error rates were excessive, particularly for failure of two or more PVTs, with the false-positive rate increasing as a function of the total number of PVTs administered. For example, failure of two or more PVTs out of five at a per-test false-positive rate of 10% yielded a false-positive rate of 11.5%, a value slightly greater than the per-test false-positive rate, but when the total number of PVTs administered was 7, 17.5% failed two or more measures, a value notably higher than the per-test false-positive rate of 10%.

In discussing their results, Berthelson and colleagues (2013) compared their Monte Carlo simulated frequencies of failure to the clinical data reported by Pella, Hill, Shelton, Elliott, and Gouver (2012). Berthelson and colleagues reported that of the 478 Pella and colleagues’ college students with no external incentive who underwent psychoeducational evaluation utilizing 11 embedded/derived PVTs, 63% failed none, 37% failed one or more, 21% failed two or more, and three or more failures were necessary to maintain a false-positive rate <10%. Using Monte Carlo simulation, creating distributions for 11 tests, with a 0.31 intercorrelation and 6% false-positive rate (based on the PVTs administered by Pella et al.), Berthelson and colleagues reported values of 63% failing none, 37% failing one or more, 16% failing two or more, and that three or more failures were necessary to achieve false-positive rates <10%, percentages quite close to Pella and colleagues’ clinical data.

This apparently impressive demonstration of concordance between the actual clinical data of Pella and colleagues (2012) and the Monte Carlo simulation of Berthelson and colleagues (2013) is likely inaccurate, primarily due to overestimation of false-positive rates in Pella and colleagues’ no-incentive clinical sample by use of multiple, nonindependent measures from the WAIS-III and WMS-III. Review of Table 2 in Pella and colleagues shows that only 2 of the 11 PVTs were independent: the WAIS-III Processing Speed Index and WMS-III Faces I raw score. Six of the 11 included measures that were derived from the Digit Span subtest, and two of these included both Digit Span and the Vocabulary subtest (the discriminant function developed by Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner, 1995, and the Vocabulary minus Digit Span difference score reported in the same paper). Three of the 11 PVTs also included WMS-III Delayed Recognition scores (Rarely Missed Index, Auditory Delayed Recognition, and the Ord et al. Index). Moreover, Pella and colleagues reported high failure rates for the no-incentive clinical sample for Mittenberg and colleagues’ (1995) discriminant function equation and Vocabulary minus Digit Span difference score (27.3%), noting that procedures based on discrepancy measures or profile differences may be suboptimal in high functioning samples such as college students, due to increased subtest scatter. In this regard, Table 7 of Pella and colleagues shows much lower false-positive rates for the no-incentive clinical sample for all other WAIS-III and WMS-III PVTs, ranging from 0.2% for WMS Faces I, to 2.6% for Reliable Digit Span and the Ord and colleagues’ Index. Indeed, summation of the false-positive rates for these other nine PVTs totaled 10.8%, despite the fact that only two of these nine remaining PVTs were independent (WAIS-III Processing Speed Index and WMS-III Faces I raw score).

Altogether, these factors strongly suggest that the failure rates reported for the no-incentive clinical sample in Table 8 of Pella and colleagues (2012) are inflated due to the elevated false-positive rates for Mittenberg and colleagues’ (1995) discriminant function equation and Vocabulary minus Digit Span difference, and the fact that 9 of the 11 PVTs were interrelated (i.e., many...
indices shared various subtests in their derivation), including 6 dependent upon Digit Span. In other words, the failure rates reported for the Pella and colleagues’ no-incentive examinees were tallied regardless of scale overlap; these were not based on administration of 11 independent PVTs. Consequently, failure on any one of the overlapping PVTs increased the probability of failing another of the overlapping PVTs (R. D. Pella, personal communication, March 17, 2014).

Review of the data reported by Berthelson and colleagues (2013) suggests false-positive rates that are greater for their Monte Carlo-simulated data, than those reported using data on actual clinical patients, as reported for the ACS PVTs (Pearson, 2009). For example, the false-positive rate for failure of 2 or more out of 5 PVTs at a per-test false-positive rate of 15%, reported by Berthelson and colleagues, 19.4% (Table V) is significantly higher than the value of 9% from Table 3.10 in the ACS Manual ($z = 5.1, p < .0001$).

False-positive rates reported by Berthelson and colleagues (2013) also appear to be larger than the values reported for clinical patients by Victor and colleagues (2009). Data in Table 5 of Victor and colleagues show an average per-test failure rate of 14.38% for 4 PVTs, with data in Table 7 showing that 47% failed 1 PVT or more, 6% failed 2 or more PVTs, 1% failed 3 PVTs, and none failed all 4 PVTs. In contrast, the data in Table V of Berthelson and colleagues show that, at a per-test false-positive rate of 15%, 40.4% fail 1 or more, 14.6% fail 2 or more, with 4.3% failing 3 or more and 0.8% failing all 4. Although Berthelson and colleagues’ Monte Carlo-simulated data yield data comparable with Victor and colleagues (2009) for failure of 1 or more PVT (40.4% vs. 47%, respectively), the failure rates are significantly higher for failure of 2 or more PVTs, 14.6% for Berthelson and colleagues versus 6% for Victor and colleagues (2009), $z = 2.0, p = .024$.

Additionally, false-positive rates reported by Berthelson and colleagues (2013) are larger than those reported by Schroeder and Marshall (2011), who studied PVT failure rates in psychotic and nonpsychotic psychiatric patients on seven PVTs, with the per-test false-positive rate set at 10% or less. Failure on any of these seven PVTs was rare in both the psychotic and nonpsychotic samples (74% and 78%, respectively, failed zero PVTs). Of particular interest, only 7% of the psychotic and 5% of the nonpsychotic patients either failed two or more PVTs or failed one PVT at worse-than-chance level of performance. These proportions, in two independent samples, are both significantly lower than the value of 17.5% reported in Table IV of Berthelson and colleagues (2013), with a $z$ score of 2.8, $p = .0024$ associated with the PVT failure rate of 7%, and a $z$ score of 4.4, $p < .0001$ associated with the 5% failure rate.

Finally, Davis and Millis (2014) have conducted a direct test of Berthelson and colleagues’ finding of an increase in false-positive rate associated with an increase in the number of PVTs administered. In a sample of 158 outpatient physiatry referrals, not screened for medico-legal involvement, there was a small nonsignificant correlation, $r_s = .13, p = .10$, between the number of PVTs failed and the number administered. Other significant bivariate findings included an inverse relationship between educational level and number of PVTs failed, inverse correlation between TBI severity and number of PVTs failed, inverse relationship between functional independence and number of PVT failures, and greater number of PVT failures in those seen in a medicolegal context. Multivariable analysis using a negative binomial regression model showed that the dependent variable, number of PVTs failed, was predicted by educational and functional status but not by the number of PVT administered, or by medicolegal context. In a subsample of 87 neurologic cases not in litigation, false-positive rates were lower than Berthelson and colleagues’ predicted rates in subjects administered six, seven, or eight PVTs using a two-PVT cut-off. Davis and Millis concluded that their data mitigated concerns that increased PVT failure was a necessary outcome of increased PVT administration. Moreover, they commented that PVT score distributions, particularly for free-standing tests, are typically skewed with a restricted range of scores in clinical cases, which questions the assumption of multivariate normality, a key requirement of the Monte Carlo simulation employed by Berthelson and colleagues.

The above review suggests that the false-positive rates based on the Monte Carlo simulations reported by Berthelson and colleagues (2013) are overestimated, in comparison to false-positive rates typically found in actual clinical patient data. The degree of overestimation appears substantial, for example, for failure of two or more out of seven tests at a 10% per-test false-positive rate, the false-positive rate of 17.5% of Berthelson and colleagues is over three times the value of 5% reported in Table IV of Berthelson and colleagues (2013) for failure of 1 or more PVT (40.4% vs. 47%, respectively), the failure rates are significantly higher for failure of 2 or more PVTs, 14.6% for Berthelson and colleagues versus 6% for Victor and colleagues (2009), $z = 2.0, p = .024$.

The purpose of the current investigation is to directly compare PVT failure rates for additional clinical samples to the Monte Carlo simulations reported by Berthelson and colleagues (2013). Data from the clinical groups in two papers by Larrabee (2003, 2009) will be analyzed for false-positive rates on four PVTs and one SVT (Larrabee, 2003) and six PVTs and one SVT (Larrabee, 2003, 2009 combined), as will data from the 10% per-test false-positive rate for the five ACS effort tests (Pearson, 2009), in comparison to Berthelson and colleagues’ (2013) data reported in their Table IV for the five and seven subtest datasets in the 10% false-positive rate per-test condition. Finally, although the primary analyses are focused on false-positive rates in the nonmalingering clinical samples in Larrabee (2003, 2009), the diagnostic accuracy associated with failure on the combined set of 6 PVTs and 1 SVT will be determined for the 54 nonmalingering clinical cases versus the 41 definite (24) and probable (17) malingers.
Method

Larrabee (2003, 2009)

Subjects

The 54 clinical subjects from Larrabee (2003, 2009) included 15 who had sustained severe traumatic brain injury (TBI) (Glasgow Coma Scale or GCS of 8 or less), and 12 who had sustained a moderate TBI (GCS of 9 to 12 or GCS of 13–15 with CT evidence of contusion). Twenty-three of the moderate/severe TBI had CT/MRI findings available for review, and 21 of these were positive for structural abnormalities consequent to brain trauma. One case with a negative MRI had positive EEG findings. The moderate/severe TBI group consisted of 14 men and 13 women, with a mean age of 34.80 (SD = 16.78) and mean education of 12.56 years (SD = 2.56). The remaining 27 subjects included a mixed neurologic group comprising 13 subjects who had a variety of neurologic disorders, and a psychiatric group comprising 14 subjects who primarily suffered from mood disorder. Neurologic disorders included multiple sclerosis (1), seizure disorder (2), stroke (2), ruptured aneurysm/AVM with surgical repair (3), herpes simplex encephalitis (1), surgically resected colloid cyst of the third ventricle with postsurgical memory disorder (1), mild cognitive impairment (2; both cases presented with primary memory deficits, and later progressed to probable Alzheimer’s disease), and 1 with multiple problems (severe TBI, stroke, and alcohol abuse). Mean age of the mixed neurologic group was 49.92 (SD = 13.16), with mean education of 14.42 years (SD = 2.75). The second group of nonlitigating clinical patients included 14 patients with primary psychiatric disorders, who had been referred for neuropsychologic evaluation due to concerns regarding their cognitive function. Eleven had major depressive disorder, two had adjustment disorder with depressed mood, and one had schizoaffective disorder. Mean age of the psychiatric subjects was 45.29 (SD = 12.21), with mean education of 14.57 years (SD = 1.99). The combined group of 54 subjects included 28 men and 26 women, with an average age of 41.16 years (SD = 16.06), and average education of 13.53 (SD = 2.62).

The malingering sample included 24 subjects in litigation or workmen’s compensation actions, who met Slick and colleagues’ criteria for definite malingering (significantly worse-than-chance performance on the Portland Digit Recognition Test or PDRT, Binder & Willis, 1991), and 17 subjects in litigation or compensation-seeking actions who met Slick and colleagues’ criteria for probable malingering (performance at or below the 2nd percentile on the PDRT plus failure on at least one other free-standing PVT such as the Test of Memory Malingering, TOMM, Tombaugh, 1996). The majority of the subjects were alleging mild TBI (mTBI). None of the subjects had objective radiologic, electrophysiologic, or laboratory values that were abnormal. See Larrabee (2003) for further information regarding this sample. For the combined definite and probable malingering group, there were 24 men and 17 women. Mean age was 40.88 (SD = 11.62), and mean education was 12.01 (SD = 2.04). The malingering sample did not differ from the clinical sample on mean age (p = .925), but did differ on mean education (p = .003). Education was not correlated in any meaningful fashion with the six PVT and one SVT measures within either group, with the majority of the correlations <.30.

Procedures

The six embedded/derived PVTs and one SVT from the Larrabee (2003, 2009) investigations and respective cut-off scores included: Benton Visual Form Discrimination (VFD; Benton, Sivan, Hamsher, Varney, & Spreen, 1994) raw score of <26; Combined dominant plus nondominant Finger Tapping (FT, Heaton, Grant, & Matthews, 1991), with a cut-off score of <63; Reliable Digit Span (RDS; Greiffenstein, Baker, & Gola, 1994), with a cut-off score of <8; Continuous Visual Memory Test (CVMT, Trahan & Larrabee, 1988) PVT score of <14; Continuous Recognition Memory (CRM, Hannay, Levin, & Grossman, 1979) PVT score of <26; Wisconsin Card Sorting Test (WCST, Heaton, Chelune, Talley, Kay, & Curtiss, 1993) Failure to Maintain Set (FMS) of >1, and MMPI-2 FBS (Lees-Haley, English, & Glenn, 1991) raw score of >21. The cut scores above were determined by contrasting performance of the definite malingers with the moderate/severe TBI subjects, attempting to keep the false-positive rate at or below 10% (see Larrabee, 2003). When the moderate/severe TBI cases were combined with the neurologic and psychiatric cases used in the cross-validation in Larrabee (2003), the false-positive rates increased for certain PVTs, for example, WCST FMS increased from 12.9% to 18.5% in the combined sample of 54 subjects. The average per-test false-positive rate for the four PVTs and one SVT in the combined sample of 54 in Larrabee (2003) was 9.6%, ranging from 3.7% (VFD, FT) to 18.5% (WCST FMS). The average per-test failure rate for the six PVTs and one SVT (combining Larrabee, 2003, 2009) was 9.0%, also ranging from 3.7% (VFD, FT) to 18.5% (WCST FMS). The average intercorrelation of all 7 scores in the combined clinical sample of 54 was 0.063.
**ACS Clinical Sample**

Subjects. For purposes of comparison to the Larrabee (2003) 4 PVT and 1 SVT data, and Berthelson and colleagues’ (2013) five PVT simulated variable model, data were included for the performance of the 371 clinical subjects on the 5 PVTs comprising the ACS (Pearson, 2009) performance validity analysis. These subjects were a subset possessing complete PVT data from a larger sample of 412 subjects (Wechsler, 2009), with the following diagnoses: moderate/severe TBI (32), right temporal lobectomy (15), left temporal lobectomy (8), schizophrenia (55), major depressive disorder (84), anxiety disorder (60), mild intellectual disability (32), autism (21), Asperger’s (35), reading disability (15), mathematics disability (22), and attention-deficit hyperactivity disorder (33). Mean age ranged from a low of 18.2 years ($SD = 2.0$) for the mathematics disability group to a high of 51.6 ($SD = 11.0$) for the major depressive disorder group. Mean education values were not reported, rather, data were reported as percent achieving each of five different education ranges (see Table 4.28, Wechsler, 2009).

Procedures. The five PVTs comprising the ACS performance validity assessment include the Word Choice Test (forced choice recognition of 50 words previously seen); Wechsler Memory Scale IV (WMS-IV, Wechsler, 2009) Logical Memory II Recognition total correct; WMS-IV Verbal Paired Associates II Recognition total correct; WMS-IV Visual Reproduction II Recognition total correct; and Wechsler Adult Intelligence Scale IV (WAIS-IV) Reliable Digit Span. Data from Table 3.9 of the ACS manual (Pearson, 2009) showing the percentage of clinical cases having various number of scores out of 5 that were less than or equal to the bottom 10% per-test performance of the general clinical sample were utilized in the present investigation. The average intercorrelation of these 5 PVTs was 0.31, as reported by Berthelson and colleagues (2013) in their Table II.

Berthelson and colleagues (2013)

Subjects and Procedures

There were no subjects, as Berthelson and colleagues utilized Monte Carlo simulation to generate test data. Sample size utilized by these authors in data simulation was 1,000,000, although values started to become invariant at 600,000 (W. Mittenberg, personal communication, October 8, 2013). Variable intercorrelation was set at 0.31. Data analyzed were for the 10% per-test failure rate, for number of PVTs out of 5 and number of PVTs out of 7, as represented in Table IV of Berthelson and colleagues (2013), determined by their Monte Carlo simulation of test data.

Results

Table 1 shows the failure rates for the number of PVTs failed out of five, for Berthelson and colleagues’ (2013) simulation data, the Larrabee (2003) data for 4 PVTs and 1 SVT, and the ACS data for 5 PVTs. Failure rates for two or more PVTs/SVTs are of interest, because these fit general criteria for probable malingering, in the context of external incentive (Slick et al., 1999) and at a base rate of malingering of 0.40, failure of two independent PVTs/SVTs, each with a sensitivity of 0.50 and specificity of 0.90, yields a PPP of malingering in the 0.90s (Larrabee, 2008). For Berthelson and colleagues’ data, 11.5% fail two or more out of five PVTs. In contrast, 5.6% of the Larrabee (2003) clinical sample fails two or more out of five PVT/SVTs, and 6% of the ACS clinical sample fails at least two of five PVTs. These proportions can be compared statistically. Contrasting Berthelson and colleagues’ figure of 11.5% with the Larrabee figure of 5.6% yields a z score of 1.4, $p = .0871$. In contrast, the comparison of 11.5% with the ACS value of 6% yields a z score of 3.3, $p = .0004$. Clearly, there is a problem of low power for the Larrabee data, given the sample size of 54, compared with the sample size of 371 for the ACS sample, and 1,000,000 for Berthelson and colleagues’ Monte Carlo-simulated data.

**Table 1.** Proportion failing PVTs and SVTs as a function of number of failures out of five at a per-test false-positive failure rate of 10%

<table>
<thead>
<tr>
<th>Number of failures</th>
<th>Berthelson$^a$</th>
<th>ACS$^b$</th>
<th>Larrabee$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.223</td>
<td>0.19</td>
<td>0.370</td>
</tr>
<tr>
<td>2</td>
<td>0.079</td>
<td>0.05</td>
<td>0.056</td>
</tr>
<tr>
<td>3</td>
<td>0.027</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes: $^a$Berthelson and colleagues, $n = 1,000,000$.

$^b$Advanced Clinical Solutions, $n = 371$.

$^c$Larrabee, $n = 54$. 
Table 2. Proportion failing PVTs and SVTs as a function of number of failures out of seven at a per-test false-positive failure rate of 10%

<table>
<thead>
<tr>
<th>Number of fails</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berthelson(^a)</td>
<td>0.236</td>
<td>0.102</td>
<td>0.044</td>
<td>0.019</td>
<td>0.007</td>
<td>0.0025</td>
<td>0.0005</td>
</tr>
<tr>
<td>Larrabee(^b)</td>
<td>0.370</td>
<td>0.074</td>
<td>0.037</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes:  
\(^a\)Berthelson and colleagues, \(n = 1,000,000\).  
\(^b\)Larrabee, \(n = 54\).

Table 3. Number of clinical and malingering cases failing PVTs and SVTs as a function of number of tests failed

<table>
<thead>
<tr>
<th>Number of fails</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Cases(^a)</td>
<td>28</td>
<td>20</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Malingering Cases(^b)</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes:  
\(^a\)Clinical cases, \(n = 54\).  
\(^b\)Malingering cases, \(n = 41\).

Table 4. Sensitivity, specificity, and total correct classification as a function of number of PVTs/SVT\(^a\) failed out of seven

<table>
<thead>
<tr>
<th>Number of failures out of six PVTs and one SVT administered</th>
<th>Clinical cases correctly identified ((n = 54))</th>
<th>Malingers correctly identified ((n = 41))</th>
<th>Total correct identification of clinical and malingering cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\geq 2) of 7 PVTs/SVT</td>
<td>48/54 (88.9%)</td>
<td>40/41 (97.6%)</td>
<td>88/95 (92.6%)</td>
</tr>
<tr>
<td>(\geq 3) of 7 PVTs/SVT</td>
<td>52/54 (96.3%)</td>
<td>36/41 (87.8%)</td>
<td>88/95 (92.6%)</td>
</tr>
<tr>
<td>(\geq 4) of 7 PVTs/SVT</td>
<td>54/54 (100%)</td>
<td>26/41 (63.4%)</td>
<td>80/95 (84.2%)</td>
</tr>
</tbody>
</table>

Notes:  
\(^a\)PVT = Performance Validity Test; SVT = Symptom Validity Test.

Table 2 shows failure rates for the number of PVT/SVTs failed out of seven for Berthelson and colleagues’ data, and the data from the Larrabee (2003, 2009) combined investigations. Again, the percent failing two or more PVTs/SVTs is of interest, and the contrast of 17.5% for Berthelson versus 11.1% for Larrabee yields a \(z\) score of 1.2, \(p = .1079\). Again, this is a function of low power given the sample size of 54 in the Larrabee data.

Table 3 shows the number of PVTs/SVTs failed out of seven for the combined clinical cases (\(n = 54\)) and the combined definite/probable malingerers (\(n = 41\)), in the combined Larrabee (2003, 2009) papers. As can be seen, there are no false positives after three PVT/SVT failures, whereas 26 definite or probable malingerers (63.4%) fail four or more PVT/SVTs.

Table 4 shows specificity, sensitivity, and total correct data as a function of number of PVT/SVT failures in the Larrabee (2003, 2009) papers. High classification rates are evident, even for the \(\geq 2\) failure condition, with 88.9% specificity, 97.6% sensitivity, and 92.6% total correctly identified. Identical total correct values are obtained for the \(\geq 3\) failure condition, with increase in specificity to 96.3%, and decline in sensitivity to 87.8%. At \(\geq 4\) PVT/SVT failures, specificity is 100% (no false positives), with 63.4% of probable/definite malingerers scoring in this range.

The six nonmalingering cases identified as false positives in the \(\geq 2\) failure condition were analyzed to determine clinical characteristics associated with false-positive identification. Five of the six cases had sustained TBI (4 severe, 1 mild complicated). For four of the five TBI cases, coma ranged from 13 days to 1 month. One severe case was a case of remote TBI as a child, seen as an adult, who had CT evidence of prior craniotomy, with bilateral frontal lobe encephalomalacia. This 44-year-old right-handed woman with 12 grades education failed the WCST (FMS = 3) and MMPI-2 (FBS = 24). A right-handed man aged 25, severe TBI (1 month coma, 2 months post-traumatic amnesia [PTA]; CT showing left basal ganglia, right temporal, and left temporoparietal hemorrhages), failed the CVMT (12), WCST (FMS = 2), and RDS (7). A 25-year-old right-handed man with 14 years education had severe TBI with 13 days of coma and 4 months of PTA, and failed the CRM (21), CVMT (11), and VFD (25). A 24-year-old right-handed man with 17 days of coma, CT scan showing subarachnoid hemorrhage, and acute hemiplegia with chronic residual weakness, failed the CRM (21) and Combined Finger Tapping (62; note right hand was 17.8 taps, left was 44.4 taps). The single case of complicated mild TBI was a 60-year-old woman with 12 years education, with brief loss of consciousness, initial Glasgow Coma Scale of 14, and CT scan showing left internal capsule/left thalamic lesion, with a second lesion in the white matter of the left parietal lobe, who failed the WCST (2 FMS) and MMPI-2 (FBS = 24). One false-positive case was from the neurologic sample, seen for evaluation following stroke, with CT scan evidence of right caudate infarct. This 61-year-old right-handed female, with 12 grades education, failed the CVMT PVT (score of 9) and the WCST FMS (2 FMS).
Perusal of the above data show that the false positives that occur share in common history of either severe TBI followed by coma, and/or complicated mild TBI, or stroke with CT scan abnormalities. One false-positive case of severe TBI clearly had an alternative explanation for failure of combined Finger Tapping, that is, medically documented acute hemiplegia, with chronic residual deficits who showed a clear pattern of lateralized impairment on Finger Tapping. Taking this factor into account, he only failed one other PVT (CRM), a performance that would not have been miss-classified as probable malingering.

Discussion

The primary finding of this study was that the PVT failure rates reported by Berthelson and colleagues (2013) based on Monte Carlo-simulated data are overestimates of the rate of PVT failure in actual clinical cases. Failure rates shown in Table 1 for two or more PVTs failed out of five, each with a per-test false-positive rate of 10%, are nearly double for Berthelson and colleagues’ data, at 11.5%, compared with the Larrabee (2003) failure rate of 5.6%, and the ACS failure rate of 6.0%. For failure of two or more out of seven PVTs/SVTs, as shown in Table 2, Berthelson and colleagues’ failure rates of 17.5% are again higher than that in the Larrabee (2003, 2009) clinical sample, at 11.1%. Moreover, the current failure rates for ≥ 2 out of 7 PVTs for both Larrabee (2003, 2009) and Berthelson and colleagues’ (2013) data are substantially higher than the rates of 7% for psychotic and 5% for nonpsychotic psychiatric patients for failure of ≥ 2 of 7 PVTs reported by Schroeder and Marshall (2011).

A small increase in false-positive rates going from four PVTs and one SVT (total of five) in Table 1 to six PVTs and one SVT (total of seven) in Table 2 is apparent, for the Larrabee (2003, 2009) data. It is noteworthy that in the Larrabee (2003) study, the false-positive rate hit zero at three or more out of five failures. In the combined seven-variable dataset in Table 2, three or more failures are associated with a 3.7% false-positive rate, not zero. Similarly, in the Larrabee (2003) five-variable dataset, failure of two or more out of five was associated with a 5.6% false-positive rate (Table 1), increasing to 11.1% for two or more out of seven (Table 2). When PVT/SVT failure rates are compared for both nonmalingering clinical cases and probable/definite malingerers (Table 3); however, once the level reaches four of seven PVT/SVT failures (zero false positives), 26 of 41 (63.4%) of the malingerers are still correctly identified, a higher level than the 21 of 41 (51.2%) identified in the Larrabee (2003) paper, when the false-positive rate became zero at three out of five failures. Finally, as shown in Table 4, the overall diagnostic classification rates are quite high, using failure of either two or three out of seven PVTs/SVT for correct identification of both nonmalingering clinical cases, and those subjects identified as probable/definite malingerers.

Thus, increasing the number of PVTs and/or SVTs administered does lead to a small increase in false-positive rate, up through three out of seven measures. Interestingly, it appears that increasing the number of PVTs leads to a slightly higher number of failures required before zero false positives occur, but the number of malingerers identified at this zero false-positive rate actually is higher than at zero false-positive rates obtained with a smaller number of PVTs administered. Since the current dataset is small, this finding certainly bears replication.

False-positive error can be further minimized by consideration of the characteristics of clinical cases classified as false positives for invalid performance. Only cases with either severe TBI, typically followed by prolonged coma and/or mild complicated cases or stroke cases with structural lesion on CT scan, failed two or more PVTs/SVTs in the present investigation. Many times failure occurred just within the range of invalid performance (e.g., RDS of seven, FMS of two). In one case, the reason for false-positive identification was clear: residual hemiparesis leading to reduced combined Finger Tapping performance. Thus, forensic cases without severe TBI or CT scan abnormalities who fail two or more out of seven PVTs/SVTs are unlikely to be misclassified as false positives. In general, the literature shows that false positives on PVTs tend to occur in persons suffering undeniably severe neurocognitive compromise, often requiring 24 h supervised care (Larrabee, 2003; Meyers & Volbrecht, 2003; Victor et al., 2009).

One can conceive of other factors that could mitigate against making false-positive errors; for example, an RDS of 7 is unlikely to be a false positive in the context of entirely normal performance on WAIS-IV Arithmetic and Letter/Number sequencing, or in the context of normal Trial 1 and List B recall on the Rey Auditory Verbal Learning Test (AVLT). A WCST FMS score of 2 is unlikely to be a false positive in an uncomplicated mild TBI case, who also performs normally on Controlled Oral Word Association, Trail Making B, Stroop, and WAIS-IV Block Design, and who also fails additional PVTs including RDS, the CRM, and CVMT. A combined Finger Tapping score of 60 is unlikely to be a false positive in the context of performance on the Grooved Pegboard falling at the 50th percentile for either hand. Failure of the Visual Form Discrimination Test at a score of 24 is unlikely to be a false positive in the context of an entirely normal WAIS-IV Perceptual Reasoning Index. Cases performing far beyond cut-off on two indicators, without evidence of moderate or severe TBI and with normal CT scan are unlikely to be false positives (e.g., RDS score of 5 and FBS raw score of 33 in an uncomplicated mild TBI). On the ACS, failure of the Word Choice Test is unlikely to be a false positive in the context of normal WMS-IV Verbal Paired Associates I and II performance.

The data from the ACS manual reported in Table 1 clearly demonstrate that the current findings are not simply the result of sample-specific data from the earlier Larrabee (2003, 2009) investigations. Indeed, the ACS false-positive rate for two or more
out of five failures, 6.0%, is essentially identical to the Larrabee (2003) rate of 5.6%. What is particularly impressive is that the ACS false-positive error rates are obtained based on clinical cases at risk for significant cognitive impairments due to neurologic, psychiatric, and/or developmental conditions, including moderate and severe TBI, temporal lobectomy, mild intellectual disability, schizophrenia, autism, anxiety disorder, and major depressive disorder.

Counting the Larrabee (2003, 2009) investigations as one clinical sample, and the ACS clinical sample as a second, there are three additional studies containing a total of four independent clinical samples, all showing lower false-positive rates in clinical cases than that reported by Berthelson and colleagues (2013) for their Monte Carlo-simulated data: Victor and colleagues (2009), Schroeder and Marshall (2011; two samples), and Davis and Millis (2014). The discrepancies in false-positive rates are particularly disproportionate in the studies using larger clinical samples, most noteworthy in the Schroeder and Marshall (2011) study wherein two or more out of seven PVTs failed at a per-test failure rate of 10% or less occurred in 7% of a sample of 104 psychotic patients, and in 5% of a sample of 178 nonpsychotic psychiatric patients, compared with the 17.5% obtained for failing two or more out of seven measures in Berthelson and colleagues' (2013) Monte Carlo-simulated data.

Lower PVT failure rate using multiple independent measures, at per-test false-positive rates of 10% or less, is apparently because these PVT data are based on performances that are atypical in both pattern and degree for what is actually seen in neuropsychologic evaluation of clinical patients with bona fide neurologic, psychiatric, and developmental disorders. It is uncommon for clinical patients to show grossly abnormal performance on tasks of recognition memory, which comprise four out of the five ACS PVT scores, and three out of seven of the PVTs in the Schroeder and Marshall (2011) investigation. It is atypical for clinical patients to fail select test items on the CRM and CVMT, and it is atypical for Victor and colleagues’ (2009) clinical subjects to perform in an abnormally poor fashion on the Rey Osterrieth Complex Figure recognition items, or on the AVLT recognition items. Consequently, PVTs are commonly performed at ceiling, with restricted range of scores, in nonmalingering clinical samples, resulting in skewed distributions of scores. For example, on the TOMM (Tombaugh, 1996) 21 aphasic patients averaged 98.7% correct on Trial 2, with 16 achieving perfect scores, 3 scoring 98%, 1 scoring 96%, and one scoring 82%. Nearly half of the 45 TBI patients reported by Tombaugh (1996), 22, had severe TBI (range of coma 1 day to 3 months), and averaged 98.3% correct on the TOMM Trial 2, with 14 achieving perfect scores, 3 scoring 98%, 2 scoring 96%, 2 scoring 94%, and 1 scoring 88%.

In contrast, Berthelson and colleagues’ (2013) data are not derived from clinically atypical performances for variables reflecting skewed distributions with restriction in range. Rather, scores for each of the simulated PVT distributions are obtained by statistical simulation generating random data, to match the standard normal distribution with a mean of 0 and standard deviation of 1.0 for each test, with the only constraints being use of scores intercorrelated at 0.31, and abnormality defined at a per-test frequency of 10% or less (Berthelson et al., 2013; Crawford et al., 2007). Berthelson and colleagues essentially could have called these scores “memory tests,” “language tests,” or “scores from a general neuropsychological battery,” rather than PVTs; indeed, these authors cited studies showing that Monte Carlo simulations did accurately capture the actual score distributions for subtests comprising various neuropsychologic batteries (Brooks & Iverson, 2010; Decker, Schneider, & Hale, 2012; Schretlen, Testa, Winicki, Pearlson, & Gordon, 2008). Berthelson and colleagues attempt at Monte Carlo simulation of PVT performance, however, is simply not specific to or characteristic of PVTs, and does not capture PVT performance variability in the same manner as it captures the variability of standard measures of neuropsychological abilities. As Davis and Millis (2014) have observed, the restriction in range and skewed distributions commonly seen with PVT performance in clinical, nonmalingering samples suggests that score distributions characteristic of PVTs violate the assumption of multivariate normality upon which the Monte Carlo simulation is based. This appears to be a reasonable explanation for the significant discrepancies between false alarm rates in actual clinical cases with bona fide neurologic, psychiatric, and developmental disorders, and the rates reported for Berthelson and colleagues’ Monte Carlo investigation.

Data are readily found in the PVT literature demonstrating how rare it is for patients with bona fide disorder to fail actual PVTs. For example, the TOMM manual (Tombaugh, 1996) shows that 20 of 21 aphasic patients examined passed the TOMM (the only failure was a patient with global aphasia), and 44/45 TBI patients, about half of whom had severe TBI, passed the test as well. Select cases performed perfectly in the context of severe disorder and obvious impairment on formal neuropsychological testing, for example, a case with colloid cyst in the third ventricle, PTSD, borderline personality disorder, alcohol abuse, hypothyroidism, and COPD, who had verbal memory learning and retention scores at or below the 1st percentile, with visual memory retention below the 1st percentile, scored perfectly on TOMM Trial 2 and Retention. A case with massive head wound due to gunshot resulting in 38 days of coma, right frontal lobectomy, seizure disorder, with additional history of prostate cancer and COPD, scored perfectly on TOMM Trial 2 and Retention. Goodrich-Hunsaker and Hopkins (2009) reported that three patients with bilateral hippocampal damage secondary to anoxia performed normally on the PVT trials of the Word Memory Test (WMT, Green, 2003). Carone, Green, and Drane (2014) also reported normal performance on the PVT trials of the WMT for two patients who had sustained destruction of the left anterior hippocampus and the parahippocampal gyrus, despite showing impaired free recall. McBride, Crighton, Wygant, & Granacher (2013) found that presence or absence of brain lesions was not associated with performance on three PVTs (TOMM, Victoria Symptom Validity Test, Slick, Hopp, Strauss, & Thompson, 1997; Letter Memory Test, Inman et al.,
In summary, the use of multiple PVTs and SVTs remains the recommended approach for evaluation of both performance and symptom validity. False-positive error rates decrease, as a function of the increase in number of PVTs and SVTs failed. The use of multiple measures does not seem to cause a significant increase in false-positive errors beyond the per-test false-positive rate, at least up through seven independent measures, in clinical cases, employing PVTs with per-test false-positive rates of 10% or less. Quantifying the PVT false-positive error rate through Monte Carlo simulation leads to overestimation of false-positive error rate, in comparison to actual clinical data. Consequently, subsequent evaluation of false-positive rates should be based on the performance of actual clinical patients, who are not in a compensation-seeking setting, or have any other appreciable external incentives.

Conflict of Interest

Dr. Larrabee receives royalties from Psychological Assessment Resources for sales of the Continuous Visual Memory Test, and receives royalties from Oxford University Press, for sales of Assessment of Malingered Neuropsychological Deficits. He also serves as an expert neuropsychologist in litigated matters.

Acknowledgements

I am grateful for the helpful comments of Scott R. Millis, Ph.D.

References


