The Dangers of Failing One or More Performance Validity Tests in Individuals Claiming Mild Traumatic Brain Injury-Related Postconcussive Symptoms

Daniel A. Proto1,2, Nicholas J. Pastorek3,4,*, Brian I. Miller1,2, Jennifer M. Romesser5, Anita H. Sim6, John F. Linck7

1Michael E. DeBakey Veterans Affairs Medical Center, Mental Health Care Line, Houston, TX, USA
2Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, TX, USA
3Michael E. DeBakey Veterans Affairs Medical Center, Rehabilitation and Extended Care Line, Houston, TX, USA
4Department of Physical Medicine and Rehabilitation, Baylor College of Medicine, Houston, TX, USA
5George E. Wahlen Veterans Affairs Medical Center, Mental Health, Salt Lake City, UT, USA
6Department of Physical Medicine and Rehabilitation, Minneapolis Veterans Affairs Health Care System, Minneapolis, MN, USA
7Oklahoma City Veterans Affairs Medical Center, Mental and Behavioral Health, Oklahoma City, OK, USA

*Corresponding author at: Michael E. DeBakey Veterans Affairs Medical Center, Rehabilitation and Extended Care Line, Houston, TX, USA.
Tel: +1-713-791-1414; fax: +1-713-794-7631.
E-mail address: nicholas.pastorek@va.gov (N. J. Pastorek).
Accepted 11 August 2014

Abstract

Evaluating performance validity is important in any neuropsychological assessment, and prior research recommends a threshold for invalid performance of two or more performance validity test (PVT) failures. However, extant findings also indicate that failing a single PVT is associated with significant changes in neuropsychological performance. The current study sought to determine if there is an appreciable difference in neuropsychological testing results between individuals failing different numbers of PVTs. In a sample of veterans with reported histories of mild traumatic brain injury (mTBI; \(N = 178\)), analyses revealed that individuals failing only one PVT performed significantly worse than individuals failing no PVTs on measures of verbal learning and memory, processing speed, and cognitive flexibility. Additionally, individuals failing one versus two PVTs significantly differed only on delayed free recall scores. The current findings suggest that failure of even one PVT should elicit consideration of performance invalidity, particularly in individuals with histories of mTBI.

Keywords: Mild TBI; Effort; Performance validity; Neuropsychological assessment

Introduction

Evaluation of the validity of cognitive testing performance has been recognized as a necessary component of the neuropsychological assessment process, with both the American Academy of Clinical Neuropsychology (Heilbronner et al., 2009) and the National Academy of Neuropsychology (Bush et al., 2005) recommending the use of validity testing in issued practice statements. This recognition of the importance of validity testing is based in part on the high prevalence of invalid responding found across various settings (Mittenberg, Patton, Canyock, & Condit, 2002) as well as the importance of practitioners remaining familiar with the base rates of such potentially frequently encountered conditions (Gouvier, Hayes, & Smiroldo, 1998; Gouvier, 1999). Additionally, scores on validity-focused measures (particularly performance validity tests, PVTs) have been found to explain more of the variance in cognitive testing results in individuals with traumatic brain injury (TBI) than do factors such as injury severity and length of loss of consciousness (LOC; Green, Rohling, Lees-Haley, & Allen, 2001; Meyers, Volbrecht, Axelrod, & Reinsch-Boothby, 2011), further highlighting the significance of assessing response bias.

Published by Oxford University Press 2014. This work is written by (a) US Government employee(s) and is in the public domain in the US.
Practitioners and researchers have proposed that effort can fluctuate throughout the course of an evaluation, and as such should be tested multiple times via multiple measures (Boone, 2007, 2009; Bordini, Chaknis, Ekman-Tuner, & Perna, 2002; Heilbronner et al., 2009). When interpreting said measures, and due to the particularly adverse consequences of misidentifying an effortful performance as invalid, neuropsychologists may feel obligated to obtain multiple indications of response bias prior to arriving at a conclusion of suboptimal effort. Such practices likely stem in part from published diagnostic criteria of overt malingering (Greiffenstein, Baker, & Gola, 1994) and malingered neurocognitive dysfunction (MND; Slick, Sherman, & Iverson, 1999) that require multiple sources of converging information, as well as recommendations that a single performance that is not far below established cutoffs may require additional corroborating evidence to be indicative of invalid responding (Bush et al., 2005). In follow-up evaluation of the Slick and colleagues (1999) criteria for MND, Larrabee, Greiffenstein, Greven, and Bianchini (2007) argued that failure of ≥ 2 independent, well-validated performance and/or symptom validity measures should be sufficient to classify an individual as probable MND, and failure of ≥ 3 such instruments is statistically equivalent to definite MND, providing support for a criteria of more than one effort test failure. Larrabee (2008) then later demonstrated that at varying base rates of malingering (ranging from 0.10 to 0.90), the post-test probabilities and post-test odds of definite MND varied much more with failure of a single PVT than with failure of two or three PVTs, again supporting a cutoff of ≥ 2 PVT failures. Prior research by Meyers and Volbrecht (2003) came to similar conclusions, directly suggesting that ≥ 2 PVT failures be used as an indicator of assessment invalidity owing to 0% of non-litigant and non-institutionalized participants performing in this range (i.e., 100% specificity). Finally, Victor, Boone, Serpa, Buehler, and Ziegler (2009) found in their sample that a pairwise comparison of any two validity test measures produced a higher hit rate in classifying non-credible individuals than any single-test alone or any combination of three failed PVTs, indicating that two validity failures might represent a “sweet spot” of sorts. However, it must be noted that Victor and colleagues (2009) used failure of ≥ 2 validity measures as a means of defining their non-credible group, which may have inflated the accuracy of the pairwise PVT criteria relative to the ≥ 1 and ≥ 3 PVT criteria.

Some clinicians and researchers have drawn attention to the potential for heightened false-positive identifications of suboptimal effort when including multiple validity measures in the course of a single evaluation (e.g., Berthelson, Mulchand, Odland, Miller, & Mittenberg, 2013), this concern has generally not been supported by empirical findings. Indeed, data have shown that when using cutoff scores that maintain adequate specificity (e.g., ≥ 0.90), the use of multiple validity measures raises overall sensitivity while not significantly reducing their specificity; in fact, such practice may even reduce overall false-positive rates and/or improve overall classification accuracy relative to that of the individual measures themselves (Greve, Bianchini, & Brewer, 2013; Inman & Berry, 2002; Iverson & Franzen, 1996; Larrabee, 2003, 2008, 2014; Meyers & Volbrecht, 2003; Meyers et al., 2014; Orey, Cragar, & Berry, 2000; Victor et al., 2009). Larrabee (2003), for example, showed that sensitivity for identifying invalid performances increased from a single-test average of 0.534 to a battery-wide 0.875 along with a concomitant increase in specificity from a single-test average of 0.907 to a battery-wide 0.944 when requiring ≥ 2 validity test failures. Victor and colleagues (2009) found similar sensitivity (0.838) and specificity (0.939) values in their sample when requiring ≥ 2 validity test failures, and their logistic regression results were most accurate when including scores from four validity measures rather than from two or three. Greve and colleagues (2008) obtained even higher specificity rates of (96 – 98%) in a sample of chronic pain and TBI patients when administering multiple PVTs and establishing thresholds of two or three below cutoff performances. Davis and Millis (2014) found that the number of PVTs administered was not predictive of the number of PVTs failed, and that in general, false-positive rates in their sample were lower than statistical predictions in individuals administered six to eight PVTs. Finally, Larrabee (2014) demonstrated that false-positive rates for PVTs derived via Monte Carlo simulations, such as those used by Berthelson and colleagues (2013), are overestimated relative to data from clinical samples and that the use of multiple validity tests does not significantly increase the type I error rates beyond those of the individual measures. Thus, failure of two or more well-validated instruments can be viewed as a reasonable threshold for identifying suboptimal effort, and evidence indicates that such a threshold should not result in high false-positive rates. Research methodologies have subsequently adopted or suggested this criterion as a means of identifying effort failure groups and/or defining poor effort (e.g., Lange, Pancholi, Bhagwat, Anderson-Barnes, & French, 2012; Van Dyke, Millis, Axelrod, & Hanks, 2013; Webb, Batchelor, Meares, Taylor, & Marsh, 2012), further indicating its increasing commonality, although little work has focused on how to classify individuals failing only one PVT. Such individuals are typically either included in the passing/credible group (e.g., Victor et al., 2009) or excluded from the study, perhaps unless falling below statistical chance on a forced-choice measure (e.g., Lange et al., 2012; Webb et al., 2012).

A small handful of studies have shown that failure of even a single validity instrument is related to significant differences in cognitive testing results relative to those obtained by individuals who did not perform below cutoff on any such measures. Green and colleagues (2001) found in their large, mixed clinical sample that participants failing one of either the Word Memory Test (WMT; Green, 2003; Green, Allen, & Astner, 1996) or Computerized Assessment of Response Bias (CARB; Allen, Conder, Green, Cox, 1997) obtained an average overall test battery mean (OTBM) that was 0.94 SD lower than participants who passed both the WMT and CARB. Relatedly, Demakis and colleagues (2008) examined the effects of both PVT and SVT
performance on neuropsychological and psychological test scores in a medico-legal sample, and found that the OTBM for cognitive measures decreased with each successive PVT failure. While no effect sizes were reported for the differences, the authors found that individuals failing no PVTs performed significantly better on OTBM and in each individual domain than did individuals failing only one PVT. Greve and colleagues (2008) showed that when adjusting cut scores on individual PVTs to allow for very low false-positive rates (i.e., ~2%), overall sensitivity in identifying individuals meeting criteria for malingering pain-related disability (Bianchini, Greve, & Glynn, 2005) naturally increased from 48% to 69% when requiring ≥1 rather than ≥2 PVT failures as a marker of response bias while overall specificity decreased by only 3% (from 98% to 95%). As these studies included mixed clinical samples, it is perhaps possible that some of the OTBM decreases and associated effect sizes may be attributable to genuine cognitive problems associated with false-positive errors. However, Meyers and colleagues (2011) found comparable results in a physician or attorney/case manager-referred sample of individuals with very mild TBI (e.g., ≤5 min LOC), and thus with essentially no expectation of widespread sizable and persisting changes in neuropsychological functioning. These authors’ data revealed a significant difference between individuals who failed none of nine embedded PVTs on the Meyers Neuropsychological Battery (Meyers & Rohling, 2004) versus those who failed only one. The effect size associated with failure of only one PVT was large (Cohen’s $d = 1.32$), and was also greater than that obtained when comparing individuals who failed one versus two PVTs (Cohen’s $d = 0.76$).

To date, then, despite recommendations that two or more PVTs should be failed before a cognitive profile is considered invalid, evidence exists that failure of only one among a group of PVTs is associated with significant declines in cognitive test performance. The question then remains as to how research and clinical data should be interpreted in the context of a single PVT failure, especially in samples where neurological burden is known to be very low. While much prior research has examined the classification accuracy of PVTs and PVT cutoff criteria (e.g., ≥2 failed) in various samples, the number of studies examining the effect sizes of changes in cognitive testing performances associated with number of PVT failures is relatively limited. Additionally, no prior studies were found that evaluated performance on multiple stand-alone and embedded PVTs in a veteran sample with a history of mild TBI (mTBI). And while a small handful of articles have discussed the effects of the number of PVT failures on neuropsychological OTBM and in individual cognitive domains, no identified studies have combined all of the aforementioned facets together to examine the effect sizes of performance differences across individual cognitive tests stratified by number of PVTs failed. The purpose of the current study was to evaluate these factors—that is, to determine the effect sizes of performance differences in multiple cognitive measures associated with increasing numbers of PVTs failed in a sample of veterans with a self-reported history of mild TBI. It was hypothesized based on existing research that just one PVT failure would be associated with widespread, moderate-to-large effect sizes across all evaluated cognitive measures while failure of subsequent PVTs would result in additional, but less substantive decline in neuropsychological test performance.

**Methods**

**Participants**

The initial study sample consisted of 229 Operation Enduring Freedom/Operation Iraqi Freedom/Operation New Dawn (OEF/OIF/OND) veterans consecutively referred for follow-up neuropsychological evaluation at four VA hospitals and medical centers located in the Western, Southern, and Midwestern United States. Automatic referrals for further assessment of these veterans were generated when, during a comprehensive and ongoing nationwide screening process, they reported suffering a potential deployment-related head injury that resulted in altered mental status and postconcussive symptoms at the time of the injury along with current persisting postconcussive symptoms (U.S. Government Accountability Office, 2008). Following receipt of this referral, veterans were evaluated by each facility’s polytrauma team and a portion was subsequently sent for follow-up neuropsychological evaluation based on perceived clinical need. All veterans presenting for neuropsychological evaluation received a standard assessment battery. This study was approved by the Institutional Review Boards of all participating entities and informed consent was obtained prior to the evaluation. All testing was performed by licensed psychologists or trainees under the direct supervision of licensed psychologists, and occurred between May 2011 and December 2013.

Demographic and injury characteristic information was obtained via semi-structured clinical interview. For the purposes of this study and consistent with American Congress of Rehabilitation Medicine (ACRM) criteria (ACRM, 1993), a mTBI was defined as a head injury resulting in reported altered mental status (e.g., disorientation, confusion), LOC lasting approximately ≤30 min, and/or post-traumatic amnesia lasting ≤24 h. Veterans’ data were excluded from the study if the veterans denied any history of head injury (before, during, and/or after deployment), if they endorsed characteristics for any prior injury more severe than an mTBI (i.e., LOC lasting >30 min and/or post-traumatic amnesia lasting >24 h), if they reported any history of psychotic disorder, or if they were not administered all of the PVT measures selected for these analyses. Additionally, individuals whose most recent head injury occurred within ~3 months of their evaluation were excluded based on data, indicating that objective cognitive
deficits related to mTBI typically do not persist beyond 1 – 3 months post-injury (Carroll et al., 2004). Thus, all individuals included in the current study were at least ~3 months removed from their most recent mTBI, with time since injury across the entire sample ranging from 116 days to ~18 years.

Of the 229 veterans included in the original sample, 14 were missing data on one or more of the six evaluated PVTs, 31 reported injury characteristics that were not consistent with an mTBI, four reported no history of head injury, six endorsed a history of psychiatric disorder during interview, and two reported experiencing at least one mTBI within ~3 months of their evaluation. Due to there being some overlap between these exclusionary group members (e.g., a veteran might be missing one of the six PVTs and also have reported a history of psychiatric disorder), a total of 51 individuals were excluded from the study analyses, resulting in a final sample size of 178 participants. For the individuals who were not administered all PVTs, there was no systematic reason for doing so (e.g., additional PVTs were not systematically abandoned if the participant failed a PVT early in the testing process). Chi-square analyses and one-way analysis of variance (ANOVA) indicated that the excluded veterans (mean age = 30.0 years, $SD = 5.67$) were significantly younger than the non-excluded veterans (mean age = 32.6 years, $SD = 7.69$; $F(1, 227) = 5.02, p < .05$) while ethnicity (Caucasian vs. non-Caucasian), years of education, and gender did not significantly differ between these groups. Of the 178 included veterans, 167 were male (94%), while self-reported ethnicity was as follows: 23 African American (13%), 120 Caucasian (67%), 28 Hispanic/Latino (16%), 5 Asian American (3%), and 2 other (1%). Average level of education for the sample was 13.2 years ($SD = 1.72$ years), and ranged from 8 to 18 years. No significant differences on any of the listed demographic factors were identified between the different PVT failure level groups ($p > .05$ in all cases). Information on injury characteristics for the entire sample, as well as demographic and injury characteristics stratified by PVT failure group, are provided in Table 1. Information regarding comorbid medical and/or psychiatric diagnoses beyond mTBI and the study exclusion criteria (i.e., psychotic disorder) was not systematically assessed and recorded. All participants were seen on an outpatient basis.

**Measures**

**Clinical Interview:** The semi-structured clinical interview used in this study was a modified form of the structured clinical interview used to evaluate all veterans in the VA Polytrauma Network System of Care (Belanger, Uomoto, & Vanderploeg, 2009), and was adopted by all sites participating in this project. The interview collected a variety of information in a standardized format, including demographic and background information such as age, ethnicity/race, and education, as well as the numbers, types, dates, and characteristics of deployment, pre-deployment, and post-deployment head injuries.

| Demographic variables and injury characteristics for PVT failure level groups |
|-----------------------------|----------------|----------------|----------------|----------------|
|                          | Entire sample ($n = 178$) | 0 PVT fails ($n = 61$) | 1 PVT fail ($n = 63$) | 2 PVT fails ($n = 25$) | >3 PVT fails ($n = 29$) |
| Age (years), mean ($SD$)   | 32.6 (7.69)    | 31.7 (7.36)    | 32.2 (7.07)    | 33.7 (7.43)    | 34.5 (9.63)    |
| Education (years), mean ($SD$) | 13.2 (1.72)    | 13.4 (1.76)    | 13.1 (1.57)    | 13.3 (2.01)    | 12.9 (1.68)    |
| Gender                     | 167 males      | 57 males       | 61 males       | 22 males       | 29 males       |
| Ethnicity                  | 23 black       | 6 black        | 10 black       | 3 black        | 4 black        |
| Number of reported mTBIs, mean ($SD$) | 2.6 (2.57)    | 2.6 (2.58)    | 2.4 (2.27)    | 2.1 (1.75)    | 3.5 (3.52)    |
| Deployment                 | 2.0 (1.86)     | 2.3 (2.35)     | 1.5 (0.76)     | 1.8 (1.33)     | 1.8 (1.92)     |
| Post-deployment            | 1.6 (2.43)     | 0.80 (0.45)    | 4.0 (5.20)     | 1.3 (5.8)      | 1.0 (0.00)     |
| Loss of consciousness in minutes for most serious injury, mean ($SD$) | 3.6 (5.59)    | 2.8 (3.85)    | 2.9 (4.22)    | 5.2 (8.83)    | 5.5 (8.03)    |
| Post-traumatic amnesia in minutes for most serious injury, mean ($SD$) | 6.6 (6.56)    | 8.2 (7.30)    | 7.6 (7.70)    | 3.4 (3.21)    | 2.0 (n/a)a     |
| Deployment                 | 7.4 (9.81)     | 3.0 (2.83)     | 5.0 (4.24)     | 3.5 (2.12)     | 29.0 (n/a)a    |

**Note:** PVT = Performance Validity Test; mTBI = mild traumatic brain injury.

aCell had an $n = 1$ so no $SD$ was calculated.

aCell had no individuals reporting post-traumatic amnesia.
Cognitive functioning: All veterans in the sample were administered a standardized battery of neuropsychological assessment instruments. From these measures, the following scales/indices and scores were included in the present study: the Processing Speed Index (PSI) of the Wechsler Adult Intelligence Scale-IV (WAIS-IV; Wechsler, 2008); Trials 1–5 raw sum (Total Learning, or TL) and long-delay free recall (LD) raw score from the California Verbal Learning Test-II (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000); raw sum of Trials 1 and 2 of the Paced Auditory Serial Addition Task (PASAT; Gronwall & Sampson, 1974, as reported in Diehr, Heaton, Miller, & Grant, 1998); raw time to completion for the Trail Making Test parts A and B (TMT A and B; Reitan & Wolfson, 1985); and raw total score for Letter Fluency (FAS; Gladsjo et al., 1999). Tests or subtests from which embedded PVTs were derived (e.g., the Digit Span subtest from the WAIS-IV) were purposefully excluded as sources of cognitive functioning variables.

Performance validity: The present study included six stand-alone and embedded performance validity measures—the WMT, Test of Memory Malingering (TOMM; Tombaugh, 1996), Rey Memory for Fifteen Items Test (MFIT; Rey, 1964) with recognition trial, Reliable Digit Span (RDS; Greiffenstein, Baker, & Gola, 1994), CVLT-II Forced Choice (CVLT-II FC), and Wisconsin Card Sorting Test (Heaton, Chelune, Talley, Kay, & Curtiss, 1993) Failure to Maintain Set (FTMS). Recommended manual-based cut scores from primary scales/trials were used for the WMT and TOMM. Cutoff values for the MFIT free recall and recognition combination score (Boone, Salazar, Lu, Warner-Chacon, & Razani, 2002), RDS (Greve, Bianchini, & Brewer, 2013; Meyers & Volbrecht, 2003), CVLT-II FC raw score (Root, Robbins, Chang, & Van Gord, 2006), and raw number of FTMS errors (Greve & Bianchini, 2007; Greve, Heinly, Bianchini, & Love, 2009) were based on published values aimed at maximizing specificity.

Procedure

Subsequent to providing informed consent, veterans participated in the clinical interview and neuropsychological testing. Given the clinical context of the evaluations, environmental and/or patient-related factors sometimes necessitated that testing battery characteristics be altered (e.g., order changed, certain tests or subtests removed, etc.). As agreed upon by all the principal investigators, however, the WMT was always the first objective cognitive measure administered in the battery. Thus, while most participants completed all measures, and while all included participants received all six PVTs, some participants did not receive all of the evaluated neuropsychological measures.

In determining failure on the WMT, the lowest of the Immediate Recall, Delayed Recall, and Consistency scales was used for each participant. Trial 2 of the TOMM was utilized as the determining score on that measure (TOMM Retention trial was not part of the standard battery administered across all participating sites). For the remaining measures, as previously mentioned, published cutoff criteria that maximized specificity while not greatly sacrificing sensitivity were used. Participants were initially assigned to one of the following seven mutually exclusive groups based on the number of PVT failures exhibited: 0 failures (n = 61), 1 failure (n = 63), 2 failures (n = 25), 3 failures (n = 14), 4 failures (n = 8), 5 failures (n = 7), and 6 failures (n = 0). Owing to the relatively small numbers of individuals in the final four PVT failure levels, these groups were collapsed into a single group representing those individuals with ≥3 PVT failures to allow for a sample size consistent with the other PVT failure groups.

Analyses

ANOVA was used to examine between-group differences across the four levels of the independent variable (i.e., 0, 1, 2, and ≥3 PVTs failed) for each of the dependent variables (cognitive testing performances). Post hoc analyses consisting of Fisher’s protected least significant difference tests were then run when significant main effects were found. Additionally, effect sizes (Cohen’s d) of the various significant and non-significant between-group differences on cognitive testing performances for each of the four PVT failure levels were calculated based on group means, SDs, and sample sizes. Participants with missing cognitive data were removed from the various analyses in list-wise fashion, resulting in slight variations in sample sizes across the ANOVAs; however, the number of cases removed in this fashion was small relative to the overall sample size. Of the 178 participants included in the study, all 178 had available data for the WAIS-IV PSI, CVLT-II Total Recall, and CVLT-II Long Delay; four individuals did not have data available for the PASAT; and one individual each did not have data available for Trails A, Trails B, and FAS. To evaluate the inter-relation of the various PVTs with one another, bivariate correlations were run to determine the Pearson correlation coefficients between each of the validity measures.

Results

The rates of failure across the various PVTs, stratified by participant group, are available in Table 2. The largest number of individuals (overall and across all participant groups) failed the WMT, followed by CVLT-II FC, RDS, TOMM, FTMS, and finally Rey 15-item Test. Interestingly, no individuals failed only the TOMM or CVLT-II FC; that is, failure of one of these
measures always resulted in failure of at least one other PVT. The results of the bivariate correlation analyses between the various PVTs are available in Table 3. Per these data, beyond the relationships of the three WMT scales with one another, the various PVTs generally exhibited small-to-moderate correlations. When examining only those individuals passing all PVTs, the correlations were much smaller, ranging from $-0.21$ to $0.3$ when not including the WMT subscale inter-correlations, and were significant only in the cases of RDS/WMT Consistency, CVLT-II FC/FTMS, and TOMM/WMT Delayed Recall. Conversely, the correlation coefficients were larger and much more similar to those listed in Table 3 when evaluating only individuals who failed at least one PVT, with values ranging from $-0.75$ to $0.61$ (again ignoring the WMT subscale inter-correlations) and that were significant in the majority of cases.

Regarding the between-groups neuropsychological data comparisons, one-way ANOVA revealed that across the four levels of PVT failure, the main effects were significant ($p < .05$) for all cognitive testing variables except for FAS ($p = .06$). As such, no post hoc analyses were conducted on FAS and it is thus excluded from further discussion in this section. As both TMT A and TMT B data violated the homogeneity of variance assumption, post hoc analyses on these variables were completed with Dunnett’s T3 tests. The results of the ANOVA as well as the average performances obtained by the various PVT failure groups are presented in Table 4.

Table 5 shows the various between-groups Cohen’s $d$ effect size calculations across the neuropsychological measures, as well as the results of the ANOVA post hoc analyses (i.e., which between-groups comparisons were statistically significant and which, as previously mentioned, were not conducted for FAS). As can be seen, the zero PVT failures and one PVT failure groups differed significantly on all of the evaluated cognitive testing variables except TMT A time and PASAT combined score; the latter difference approached, but did not reach, significance ($p = .054$). The zero PVT failures and two PVT failures groups significantly differed on WAIS PSI, CVLT-II TL, and CVLT-II LD, but not on TMT A time, TMT B time, or PASAT combined score. The zero PVT failures and $\geq 3$ PVT failures groups significantly differed on all cognitive testing variables. The one PVT failure and two PVT failures groups significantly differed only on CVLT-II LD. The one PVT failure and three PVT failures groups significantly differed on all cognitive variables other than PASAT and TMT B time. Finally, the two PVT failure and three PVT failure groups significantly differed only on WAIS-IV PSI. Regarding effect size findings, when between-groups differences on cognitive testing scores were statistically significant among the various levels of the PVT failures variable, the associated effect size values ranged from medium to large per published interpretation guidelines (Cohen, 1988). For non-significant between-groups differences, the associated effect sizes were generally small.

### Discussion

The present study sought to evaluate the differences in cognitive testing performances between groups of individuals with a reported history of mild TBI exhibiting various levels of PVT failure. To that end, mutually exclusive groups denoting the

---

**Table 2.** PVT failure rates by group

<table>
<thead>
<tr>
<th>PVT failure group</th>
<th>Overall rate ($N = 178$)</th>
<th>1 PVT fail ($N = 63$)</th>
<th>2 PVT fails ($N = 25$)</th>
<th>$\geq 3$ PVT fails ($N = 29$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT$^*$</td>
<td>n = 110 (61.8%)</td>
<td>n = 56 (88.9%)</td>
<td>n = 25 (100%)</td>
<td>n = 29 (100%)</td>
</tr>
<tr>
<td>RDS$^*$</td>
<td>n = 20 (11.4%)</td>
<td>n = 3 (4.8%)</td>
<td>n = 6 (24%)</td>
<td>n = 11 (37.9%)</td>
</tr>
<tr>
<td>CVLT-II FC$^c$</td>
<td>n = 32 (18.0%)</td>
<td>n = 0 (0%)</td>
<td>n = 7 (28%)</td>
<td>n = 25 (86.2%)</td>
</tr>
<tr>
<td>FTMS$^a$</td>
<td>n = 16 (9.0%)</td>
<td>n = 2 (3.2%)</td>
<td>n = 6 (24%)</td>
<td>n = 8 (27.6%)</td>
</tr>
<tr>
<td>Rey FIT$^e$</td>
<td>n = 15 (8.4%)</td>
<td>n = 2 (3.2%)</td>
<td>n = 2 (8%)</td>
<td>n = 11 (37.9%)</td>
</tr>
<tr>
<td>TOMM$^*$</td>
<td>n = 29 (16.3%)</td>
<td>n = 0 (0%)</td>
<td>n = 4 (16%)</td>
<td>n = 25 (86.2%)</td>
</tr>
</tbody>
</table>

**Note:** Cut scores for PVTs: $^*P$ manual, $^{b, c}14$, $^{d, e}2$, and $^{*, c}20$.

**Table 3.** Correlations between administered PVTs across the entire sample

<table>
<thead>
<tr>
<th></th>
<th>WMT IR</th>
<th>WMT DR</th>
<th>WMT CS</th>
<th>RDS</th>
<th>CVLT-II FC</th>
<th>FTMS</th>
<th>Rey FIT</th>
<th>TOMM (Trial 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT IR</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMT DR</td>
<td>.91**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMT CS</td>
<td>.85**</td>
<td>.88**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDS</td>
<td>.39**</td>
<td>.43**</td>
<td>.41**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVLT-II FC</td>
<td>.51**</td>
<td>.52**</td>
<td>.44**</td>
<td>.36**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTMS</td>
<td>-.15</td>
<td>-.19*</td>
<td>-.17*</td>
<td>-.11</td>
<td>-.12</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rey FIT</td>
<td>.41**</td>
<td>.43**</td>
<td>.34**</td>
<td>.29**</td>
<td>.47**</td>
<td>-.09</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>TOMM (Trial 2)</td>
<td>.61**</td>
<td>.63**</td>
<td>.50**</td>
<td>.35**</td>
<td>.63**</td>
<td>-.13</td>
<td>.52**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Note:** $^*p < .05$; $^{**}p < .01$. 

---

*Downloaded from https://academic.oup.com/acn/article-abstract/29/7/614/2726852 by guest on 14 December 2018*
The current results generally, although not universally, supported our hypotheses. Individuals failing one PVT performed significantly worse on measures of processing speed (WAIS-IV PSI), learning and memory (CVLT-II TL and CVLT-II DR), and cognitive flexibility and set-shifting (TMT B time) than did individuals failing no PVTs. However, the associated effect sizes were generally medium rather than medium to large, and three tests—TMT A time, FAS, and PASAT—did not exhibit significant differences between the zero PVT Failures and one PVT Failure groups. The reduced effect sizes of the currently observed differences relative to prior studies (e.g., Green et al., 2001; Meyers et al., 2011) may be due in part to the fact that individuals were not selected based on active pursuit of secondary gain and the varying levels of sensitivity of the studies’ selected validity measures. Additionally, in those previous studies in which fewer PVTs were administered, the likelihood of failing multiple PVTs may have been reduced, consequently resulting in an increased effect size associated with only one PVT failure. Also, as mentioned, identifiable performance decrements (e.g., results falling significantly outside the average range) were not observed across all cognitive domains. These findings may be due to individuals failing one or more PVTs having difficulty maintaining a consistent level of suboptimal performance across multiple measures and domains, and therefore might indicate that in a single assessment battery, not all individual data points may theoretically be invalid. Additionally, test administration order may have influenced on which measures individuals performed poorly, with participants using perceived level of performance on early PVTs and cognitive tests alike to tailor efforts on subsequent measures. Finally, the operational paradigm of the individual PVTs may relate to the particular cognitive domains affected; that is, failure of memory-based PVTs may be associated with poorer performance on objective memory measures. This idea would be tentatively supported by current findings in the one PVT Fail group, where the largest proportion of individuals failed a memory-oriented PVT (i.e., the WMT) and also exhibited the largest effect sizes relative to controls on memory-related cognitive domains (i.e., CVLT-II Total Learning and Long Delay).

However, it is essentially impossible to determine the exact cognitive domains affected, and to thereby parcel out which neuro-psychological data may or may not be valid, as even results that were not significantly below average and/or significantly different to controls may differ. Further research in this area is needed to elucidate the etiology of these differences in performance and to identify specific cognitive domains that are most affected by failure on the Performance Validity Test. For example, it is possible that individuals who fail the Performance Validity Test may have difficulties with sustained attention, focus, or motivation, which could impact performance on subsequent cognitive tests.

In conclusion, the results of the current study support the hypothesis that individuals failing one or more PVTs may exhibit significant and widespread neuropsychological performance decrements, while failure of subsequent PVTs would result in additional, but less substantive declines. However, more research is needed to determine the specific cognitive domains affected and the mechanisms underlying these performance decrements.
from the zero PVT Fails group still cannot be assumed to represent optimal effort. Regardless, the data indicate that on a variety of neuropsychological instruments, even a single PVT failure is associated with an approximately one-half SD decrease in performance in a sample whose injury history (i.e., mTBI) does not account for broadly lower than expected performance on standardized testing.

Unsurprisingly, the \( \geq 3 \) PVT failures group performed significantly worse on all neuropsychological measures (other than FAS, for which no post hoc analyses were run owing to the non-significant ANOVA main effect) than the zero PVT failures group, with the associated effect sizes predominantly being large. Additionally, and as was initially hypothesized, groups with greater numbers of PVT failures generally exhibited a step-wise decline in cognitive testing data with each subsequent additional failed validity test, exhibiting increasingly poor performances relative to groups with fewer PVT failures. Of potentially greater interest, however, these findings were not true across all neuropsychological measures and for all PVT failure levels. Of particular note, relative to individuals who failed a single PVT, the two PVT Failures group significantly differed only in their delayed auditory verbal memory recall scores (i.e., CVLT-II LD). That is, failure of a second PVT resulted in significantly worse performance only on a measure of delayed free recall ability; on measures of initial learning, processing speed, executive functioning, and sustained auditory attention, further significant decline was not observed.

Combined with prior research indicating very high specificity values associated with failure of \( \geq 3 \) PVTs (e.g., Larrabee, 2003; Greve et al., 2013; Victor et al., 2009), the current findings suggest the possibility that different subpopulations may exist among patients exhibiting response bias. That is, individuals performing below cutoff on three or more validity measures may represent a somewhat unique population when compared with people who fail only one or two PVTs. These individuals are seemingly more willing to grossly exaggerate or feign significant cognitive deficits, or to put forth very minimal effort during testing, as exhibited by the large effect size differences relative to controls across all of the evaluated neuropsychological measures. In that way, then, these individuals may be similar to patients who perform below chance on forced-choice validity measures. Indeed, per Larrabee and colleagues (2007) discussion, these individuals are statistically, although not conceptually, commensurate with Slick and colleagues (1999) definite MND group. This idea then raises the theoretical question of whether these potentially disparate response bias populations are substantively different in other ways, such as in psychiatric, characterologic, psychosocial, and/or perhaps even biological factors.

However, of potentially greater direct clinical and research importance, the current findings suggest that the application of a blanket \( \geq 2 \) PVT cutoff as a marker of profile invalidity is potentially problematic. If failure of just one PVT results in significant decrements in neuropsychological testing performances across a variety of cognitive domains, as was demonstrated in this and prior studies (e.g., Demakis, Gervais, & Rohling, 2008; Green et al., 2001; Meyers et al., 2011), clinicians are placed in a bind if prevailing recommendations indicate that \( \geq 2 \) PVTs must be failed for data to be considered invalid. Such interpretation becomes particularly problematic when one considers that the current data show that performances in the assessed domains other than delayed free recall memory (i.e., processing speed, cognitive flexibility/set-shifting, and initial auditory verbal learning) do not statistically differ between individuals failing one PVT and those failing two PVTs. That is, in these cognitive areas, despite the potentially marked difference in the validity classification (i.e., valid vs. invalid) of these individuals based on a blanket \( \geq 2 \) PVT criteria, their objective cognitive data are remarkably similar. It stands to reason, then, that individuals failing one and two validity measures may belong to the same population, and that this population is fundamentally different from individuals failing no validity measures. While the current study provided evidence of this premise in a population with remote history of mTBI, further research is needed to determine if it also holds true in mixed clinical samples with more substantial neurological burden. Relatedly, researchers are also left with a quandary: should individuals failing one PVT be placed in “credible” groups, and thereby possibly obfuscate findings relative to non-credible groups, or be left out of the study entirely and risk exaggerating between-groups differences?

It must be noted that in evaluating the current results, multiple cautions are warranted. As the current study evaluated only clinically referred veterans with a history of mTBI, and as the sample was thus fairly homogenous in terms of education (typically high school or greater), ethnicity (majority Caucasian), age, and injury characteristics, the generalizability of our findings to other populations—particularly older individuals and/or those with other neurological conditions such as dementia, stroke, or more severe TBI—is uncertain. Also, a relatively large proportion of our sample (\( \sim 66\% \)) performed below recommended cut scores on at least one PVT, although this degree of PVT failure is comparable with rates seen in other studies evaluating similar populations (e.g., Armistead-Jehle, 2010; Whitney, Davis, Shepard, & Herman, 2008; Young, Kearns, & Roper, 2011; Young, Baughman, & Roper, 2012). Nonetheless, base rates play a considerable role in clinical decision-making, and can of course significantly affect the positive and negative predictive values of PVTs (e.g., Victor et al., 2009), resulting in potentially differing preferences for the number of failed PVTs to be used as a marker of data invalidity. Additionally, comorbid medical and psychiatric diagnostic information beyond exclusionary criteria (i.e., history of psychotic disorder) was not available for this sample. However, all participants were seen on an outpatient basis and were able to complete all or nearly all of the neuropsychological battery, and thus are likely comparable with individuals with a self-reported history of mTBI seen in various other settings. The
current study also used a combination of embedded and stand-alone PVTs while not including any measures of self-report validity (i.e., symptom validity). However, as the crux of the study revolved around influences of suboptimal effort on cognitive testing data rather than on diagnostic accuracy such as via the Slick and colleagues (1999) criteria, symptom validity measures were purposefully excluded. And the inclusion of both embedded and stand-alone measures may allow for greater generalizability to various settings in which combinations of such measures may be used. Finally, as is often the case in populations with a history of mTBI, injury characteristic data were informed solely by self-report, and thus may not entirely reflect true injury characteristics. Conversely, strengths of the study include adherence to ACRM criteria for defining a history of mTBI, the use of multiple stand-alone and embedded measures of performance validity, inclusion of a variety of commonly used neuropsychological measures, recruitment across a number of different hospitals, and evaluation of varying levels of PVT failure rates (e.g., not only those who passed all measures versus those who failed two or more).

As the current study indicates, in persons experiencing a condition with low expected persisting neuropsychological burden (i.e., history of mTBI), performances of those individuals failing only one PVT are generally more similar to those failing two PVTs than zero PVTs, and performances by both of these groups are significantly different from those obtained by individuals failing no PVTs. Additionally, failure of a single PVT is associated with significant, multi-domain reductions in neuropsychological testing performance. Thus, it stands to reason that in such individuals, failure of even one PVT (rather than of ≥ 2 PVTs) should raise concerns about performance invalidity. In building on the current findings, future studies may wish to examine the effect size differences in neuropsychological test performances of a single PVT failure as relates to the sensitivity and specificity of the particular measures and/or cutoff scores utilized. It might be expected that measures or cut scores with greater sensitivity would be associated with smaller cognitive testing score changes than those with lower sensitivity, and in this way there might exist an effect size and sensitivity trade off. Additionally, the score on individual PVTs at which failure of an additional PVT is highly likely, and thus resulting in at least two failed PVTs, may also be clinically informative to explore, as would the effect sizes associated with different degrees of failure on a single PVT. Indeed, the large number of failures on the WMT relative to all other validity tests in the current study suggests that there might exist significant differences across these groups in terms of associated effects on cognitive testing data and the likelihood of failing or passing additional PVTs.

Findings such as those seen here may temper the recommendation that two or more PVTs should fall below normative cut scores before performance is considered invalid. Indeed, the results of the current study indicate that failure of a single PVT is associated with significant decrements across a variety of neuropsychological measures, and thus the data from individuals failing even one validity measure, particularly in a population with history of mTBI, must be interpreted with caution. More generally, these data indicate that it might be in practitioners’ best interests to move away from “hard and fast” rules of thumb when it comes to test data validity. Rather, test- and sample-dependent characteristics must begin to come into play via methodology such as the chaining of likelihood ratios as initially suggested by Larrabee (2008) and further supported by Meyers et al. (2014). Existing data from these and other (e.g., Larrabee, 2014) studies indicate that doing so adequately protects against false-positive errors, which was an initial driving force behind suggestions of a ≥ 2 PVT failure threshold, while also bolstering sensitivity and taking into account data such as base rates of suboptimal performance. Such methodology would also allow for practitioners to calculate the post-test probability associated with failure of even a single validity measure in the population of interest, and thereby determine the confidence with which to label significant decrements in neuropsychological test performances such as those seen in the current study as due to suboptimal effort or genuine pathology.

Statement of Support: This material is based upon work supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development; B6812C (PI- Levin HS) VA RR&D Traumatic Brain Injury Center of Excellence, Neurorehabilitation; Neurons to Networks; Minnesota Veterans Medical Research & Education Foundation. No pharmaceutical or corporate funding was used in the preparation of this manuscript or collection of data.

Conflict of Interest

None declared.

References


