



Strength in Numbers or Quality over Quantity? Examining the Importance of Criterion Measure Selection to Define Validity Groups in Performance Validity Test (PVT) Research

Jason R. Soble^{1,2} · W. Alexander Alverson³ · Jacob I. Phillips³ · Edan A. Critchfield³ · Chrystal Fullen³ · Justin J. F. O'Rourke³ · Johanna Messerly³ · Jonathan M. Highsmith³ · K. Chase Bailey^{3,4} · Troy A. Webber⁵ · Janice C. Marceaux³

Received: 2 July 2019 / Accepted: 18 December 2019 / Published online: 9 January 2020

© This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2020

Abstract

Mirroring clinical guidelines, recent Performance Validity Test (PVT) research emphasizes using ≥ 2 criterion PVTs to optimally identify validity groups when validating/cross-validating PVTs; however, even with multiple measures, the effect of *which* specific PVTs are used as criterion measures remains incompletely explored. This study investigated the accuracy of varying two-PVT combinations for establishing validity status and how adding a third PVT or applying more liberal failure cut-scores affects overall false-positive (FP)/-negative (FN) rates. Clinically referred veterans ($N = 114$; 30% clinically identified as invalid) completing a six-PVT protocol as during their evaluation were included. Concordance rates were calculated across all possible two-and three-PVT combinations at conservative and liberal cutoffs. Two-PVT combinations classified 72–91% of valid (0–4% FPs) and 17–74% of invalid (0–40% FNs) cases, and three-PVT combinations classified 67–86% of valid (0–6% FPs) and 57–97% of invalid (0–24% FNs) at conservative cutoffs. Liberal cutoffs classified 53–86% of valid (0–15% FPs) and 39–82% of invalid (0–30% FNs) cases for two-PVT combinations and 46–75% of valid (3–27% FPs) and 60–97% of invalid (0–17% FNs) cases for three-PVT combinations. Irrespective of whether a two-or three-PVT combination or conservative/liberal cutoffs were used, many valid and invalid cases failed only one PVT (3–68%). Two-PVT combinations produced high FNs and were less accurate than three-PVTs for detecting invalid cases, though variable accuracy was found within both types of combinations based on the specific PVTs in the combination. Thus, both PVT quantity *and* quality are important for accurate validity classification in research studies to ensure reliability and replicability of findings. Applying more liberal cutoffs yielded increased sensitivity, but with generally higher FPs yielding problematic specificity, particularly for three-PVT combinations.

Keywords Neuropsychology · Performance validity · Psychometrics · Assessment · Veterans

✉ Jason R. Soble
soble@uic.edu

¹ Department of Psychiatry, University of Illinois College of Medicine, 912 South Wood Street, Chicago, IL 60612, USA

² Department of Neurology, University of Illinois College of Medicine, Chicago, IL, USA

³ Psychology Service, South Texas Veterans Healthcare System, San Antonio, TX, USA

⁴ Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, TX, USA

⁵ Rehabilitation and Extended Care Line, Michael E. DeBakey VA Medical Center, Houston, TX, USA

The establishment of practice standards calling for the routine assessment of performance validity in *all* neuropsychological evaluations has resulted in a proliferation of Performance Validity Test (PVT) research (Martin et al., 2015). Researchers have pursued the development and validation of new PVTs, as well as cross-validation of existing PVTs, in diverse clinical populations. Traditionally, simulator studies and studies involving samples with high motivation to feign (e.g., personal injury litigants) were the primary manner in which the accuracy and utility of these tests were investigated. More recently, researchers have sought to validate experimental PVTs using other well-validated PVTs as the criterion for classifying sample participants into valid and invalid/noncredible groups. While some research has begun to appreciate how base rate of failure or severity of failure can be included in operationalizing criterion measures of invalid performance

(e.g., Erdodi, 2019), best practices for these validation/cross-validation research studies have yet to be fully established. As such, research is needed to determine how many and which PVTs should be used as the criterion measures of performance validity or invalidity in these research studies.

While PVT research occasionally relies on a single criterion PVT for classifying performance as valid or invalid, the implications of failing a single criterion PVT remain inconclusive (Lippa, 2018). Recently, Schroeder et al. (2019) demonstrated the limitations of using just one criterion PVT to establish validity groups and advocated for the use of ≥ 2 PVT failures for identifying criterion groups in PVT research—a recommendation that closely mirrors the standard of practice for clinical evaluations (Larrabee, 2014). Importantly, although combining criterion measures increases the specificity of either test alone, the sensitivity of two tests considered together will, by definition, be lower than either test alone (Weinstein et al., 2005). The consequence is that researchers must avoid requiring failure on too many criterion PVTs to establish invalidity, particularly those that are inappropriate for their population or have poor psychometric properties, lest sensitivity is sacrificed for an inconsequential increase in specificity. Said another way, using more PVTs is not always better.

Lippa (2018) also reviewed literature describing the effect of “chaining” PVTs that were strongly correlated and suggested that using multiple, highly correlated PVTs as criterion measures does not improve diagnostic accuracy. However, she identified that this literature was limited by samples that were not generalizable (e.g., samples that included only severely impaired or only cognitively normal subjects). To date, there is scant literature pertaining to which PVTs maximize classification accuracy when used in combination. With these considerations in mind, this study aimed to explore the implications of using different PVT combinations as criterion measures to classify validity group status, as well as investigate whether adding additional PVTs increases accuracy in establishing validity groups for PVT research studies. Finally, given emerging research examining how PVT cut-scores influence failure base rates (e.g., An et al., 2017), this study also aimed to investigate the effect of adjusting individual PVT failure cutoffs to be more conservative or liberal on overall classification accuracy.

Materials and Methods

Participants

One hundred fourteen nonconsecutive veterans clinically referred for neuropsychological services from 2015 to 2017 at a VA medical center who completed a PVT protocol consisting of the Word Memory Test (WMT; Green, 2003), Test of Memory Malingering (TOMM; Tombaugh, 1996) Trial 1,

Advanced Clinical Solutions Word Choice Test (WCT; Pearson, 2009), Dot Counting Test (DCT; Boone et al., 2002), Reliable Digit Span (RDS; Greiffenstein et al., 1994) from the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008), and Rey 15-Item Test (RFIT; Rey, 1964) as part of their clinical evaluations and consented to include their data as part of ongoing, IRB-approved database study were included. Selection of PVTs administered, and validity cutoffs interpreted, were based on a combination of factors, including frequency of use (Martin et al., 2015; Slick et al., 2004), research support, and classification accuracy. Missing data were as follows: 1 RDS, 1 WCT, and 8 RFIT (due to it being added shortly after the study began). Neuropsychological test batteries differed among patients as clinically indicated, but in addition to the six standard study PVTs, various embedded PVTs, typically within standard memory tests or global cognitive measures, were also administered. Although these embedded PVTs varied between participants and were not examined in this study, they were considered when determining validity status on clinical evaluation. Thus, in total, each participant completed the six PVTs from the study protocol and at least one additional embedded PVT (see Table 1) during their evaluation, for a total of ≥ 7 PVTs per test battery that were used to determine overall validity status at evaluation time.

Clinical classification of validity status was made at the time of each patient’s neuropsychological evaluation by the evaluating board-certified clinical neuropsychologist based on (1) failures on at least two PVTs (2 freestanding, 2 embedded, or 1 freestanding and 1 embedded) and (2) established objective criteria for identifying noncredible performance (i.e., Slick et al., 1999), which examines for marked discrepancies between test performance and accepted models of central nervous system dysfunction, observed behavior, self- or collateral reports, or documented clinical history. Given a veteran sample in which patients largely receive their healthcare in one system, neuropsychologists had access to complete medical records (e.g., histories, neuroimaging, laboratory workup, prior evaluations) and information on service-connection/disability-seeking status to incorporate into clinical decision-making regarding validity status. Based on PVT performance and Slick criteria, 80/114 (70%) were clinically classified as valid and 34/114 (30%) as invalid based on their evaluation. For all cases in this study, PVT scores and these histories were reviewed by the authors to ensure the above criteria were met. Of the 80 valid cases, 43 (54%) met formal *Diagnostic and Statistical Manual of Mental Disorder-Fifth Edition* (DSM 5; APA, 2013) criteria for a mild ($N=35$; 81%) or major ($N=8$; 18%) neurocognitive disorder, whereas 37 (46%), despite having subjective complaints (or provider concern for possible cognitive impairment), were cognitively unimpaired

Table 1 Performance Validity Tests (PVTs) and associated conservative and liberal cut-scores for test failure for the study PVT protocol and ancillary embedded measures

	Conservative cutoff(s) for test failure	Reference(s)
Study PVT		
WMT	≤ 82.5% on any primary effort subtest without the Genuine Memory Impairment Profile for those with clinical history of cognitive impairment	Green (2003) Green et al. (2011) Alverson et al. (2019)
TOMM T1	≤ 40	Denning (2012)
WCT	≤ 41	Bain & Soble (2019) Bain et al. (2019)
DCT	≥ 14 Depression; PTSD; anxiety; mild TBI ≥ 15 learning disorder ≥ 19 mild neurocognitive disorder ≥ 20 schizophrenia; TBI with positive imaging ≥ 21 mild dementia	Boone et al. (2002)
RDS	≤ 6	Schroeder et al. (2012)
RFIT	< 9	Lezak et al. (2012)
Embedded PVTs		
HVLT-R	Recognition discrimination ≤ 5 (nonamnesic disorder)	Bailey et al. (2018a)
BVMT-R	Recognition discrimination ≤ 3 (amnesic disorder) Recognition discrimination ≤ 4 (nonamnesic disorder)	Bailey et al. (2018b)
CVLT-II	Recognition discrimination ≤ 3 (amnesic disorder) Forced choice ≤ 14	Schwartz et al. (2016)
RAVLT	Forced choice ≤ 13	Poreh et al. (2016)
RBANS	Effort index > 3 (nonamnesic disorder) Effort scale < 12 (amnesic disorder)	Silverberg et al. (2007) Novitski et al. (2012)
Liberal cutoff(s) for test failure		
Reference(s)		
Study PVT		
WMT	≤ 82.5% on any primary effort subtest without the Genuine Memory Impairment Profile for those with clinical history of cognitive impairment	Green (2003) Green et al. (2011) Alverson et al. (2019)
TOMM T1	≤ 43	Rai & Erdodi (2019)
WCT	≤ 45	Erdodi et al. (2014)
DCT	≥ 14 depression; PTSD; anxiety; mild TBI ≥ 15 learning disorder ≥ 19 mild neurocognitive disorder ≥ 20 schizophrenia; TBI with positive imaging ≥ 21 mild dementia	Boone et al. (2002)
RDS	≤ 7	Greiffenstein et al. (1994)
RFIT	≤ 11	Poynter et al. (201)

PVT Performance Validity Test, *WMT* Word Memory Test, *TOMM* Test of Memory Malingering, *WCT* Advanced Clinical Solutions Word Choice Test, *DCT* Dot Counting Testing, *RDS* Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) Reliable Digit Span, *RFIT* Rey 15-Item Test, *HVLT-R* Hopkins Verbal Learning Test-Revised, *BVMT-R* Brief Visuospatial Memory Test-Revised, *CVLT-II* California Verbal Learning Test-Second Edition, *RAVLT* Rey Auditory Verbal Learning Test, *RBANS* Repeatable Battery for the Assessment of Neuropsychological Status

based on their objective evaluation results. Diagnoses for those without cognitive impairment were no diagnosis ($N=11$), posttraumatic stress (PTSD; $N=3$), mood disorder ($N=10$), anxiety disorder ($N=4$), sleep disorder ($N=4$), or other psychiatric disorder ($N=5$). Diagnoses for those with cognitive impairment in this mixed clinical sample were stroke/cerebrovascular disease ($N=15$), Alzheimer's disease/mild cognitive impairment-amnesic subtype ($N=6$), epilepsy ($N=4$), attention-deficit/

hyperactivity disorder (ADHD)/learning disorder ($N=4$), moderate-severe traumatic brain injury (TBI; $N=3$), frontotemporal degeneration ($N=2$), Parkinson's disease ($N=1$), substance-induced ($N=1$), and multiple comorbid etiologies ($N=7$). Diagnoses for the 34 invalid cases were TBI ($N=11$; 10 mild/1 moderate); primary PTSD, depression, or anxiety ($N=10$); primary PTSD or depression with medical comorbidity (e.g., remote stroke, epilepsy, tumor resection; $N=6$); and possible malingering ($N=7$).

Measures

Of note, TOMM Trial 1 (as opposed to the full TOMM) was used given evidence for good diagnostic accuracy as a briefer freestanding PVT among veterans (Webber et al., 2018a), with pooled sensitivity (77%) and specificity (92%) (Denning, 2012; Martin et al., 2019). (See Table 1 for the conservative and liberal pass/fail cut-scores for the six PVTs included in this study protocol.)

Data Analyses

Chi-square tests and analyses of variance (ANOVAs) were assessed for systematic differences in demographic characteristics between valid and invalid cases. Correlations between PVTs were examined for valid cases. Failure base rates and mean number of PVT failures were subsequently calculated for valid and invalid cases using both conservative and liberal cutoffs. Next, all possible two- and three-PVT combinations extracted from the six PVTs in the protocol were established. Concordance rates were then calculated between validity group classification and each two- and three-PVT combination to establish the accuracy of each PVT combination for identifying invalid performance. For all two- and three-PVT combinations, failure on 0 PVTs was operationalized as valid performance and failure on ≥ 2 PVTs as indicating invalid performance. All concordance analyses were conducted once using conservative cut-scores for criterion PVT failure and then repeated with liberal PVT failure cut-scores to examine and compare potential effect(s) of different criterion cut-scores on overall classification accuracy.

Results

The sample was 83% male, but otherwise rather diverse in terms of age ($M = 54.2$; $SD = 14.9$; range = 24–84), education

($M = 13.7$; $SD = 2.4$; range = 7–19), and race (46% Caucasian, 35% Hispanic, 16% African American, 3% Other), with non-significant demographic differences between valid and invalid participants: age, $F(1, 113) = 2.33$, $p = 0.13$; education, $F(1, 113) = 0.45$, $p = 0.50$; sex, $\chi^2(1, N = 114) = 0.13$, $p = 0.71$; and race, $\chi^2(3, N = 114) = 3.72$, $p = 0.29$. Means/standard deviations for the six PVTs by validity group and correlations between the six PVTs among the valid cases are included in Table 2. Failure base rates for each PVT at conservative and liberal cutoffs as well as mean number of PVTs failures by validity group are presented in Table 3 and overall PVT failure percentages are included in Fig. 1. For invalid cases, failure rates ranged from 29 to 94% at conservative and 45 to 94% at liberal cutoffs. Among valid cases, failure rates ranged from 5 to 16% at conservative and 5 to 28% at liberal cutoffs. Across both valid, $t(79) = -6.30$, $p < 0.001$, $d = 0.57$, and invalid, $t(33) = -4.74$, $p < 0.001$, $d = 0.38$, cases, the mean number of PVT failures was significantly higher with liberal versus conservative cutoffs, though a substantially larger difference and effect size was noted for valid cases relative to invalid cases.

At conservative cutoffs, when two PVTs were used, 72–91% of clinically valid cases would be correctly identified as valid (i.e., failed 0/2 PVTs) regardless of which two PVTs were used as criterion measures, and a very small percentage (i.e., 0–4%) would be incorrectly classified as invalid (i.e., false positives) by failing 2/2 PVTs. (See Tables 4 and 5 for all possible two-PVT combinations using conservative and liberal cut-scores, respectively.) In contrast, for invalid cases, correct classification (i.e., failure on 2/2 PVTs) varied widely from 17 to 74% based on the two-PVT combination used. WMT/TOMM (74%), WMT/WCT (67%), WMT/DCT (62%), and TOMM/WCT (58%) had the highest concordance, whereas RDS/RFIT (17%), TOMM/RFIT (19%), DCT/RFIT (23%), and WCT/RFIT (27%) had the lowest. Misclassification of invalid performance as valid (i.e., failure on 0/2 PVTs or false negatives) also varied considerably with few to none for all WMT

Table 2 Performance Validity Tests scores by validity group and correlations for the valid cases

	Valid <i>M</i> (<i>SD</i>)	Invalid <i>M</i> (<i>SD</i>)	<i>F</i>	ηp^2	<i>IR</i>	<i>DR</i>	<i>CNS</i>	<i>TOMM</i>	<i>WCT</i>	<i>DCT</i>	<i>RDS</i>	<i>RFIT</i>
WMT- <i>IR</i>	90.87 (10.5)	66.91 (13.7)	90.94***	0.47	–	0.82**	0.89**	0.52**	0.52**	–0.04	–0.08	0.40**
WMT- <i>DR</i>	90.68 (11.1)	63.64 (16.5)	92.79***	0.48	–	–	0.92**	0.56**	0.59**	0.01	–0.10	0.40**
WMT- <i>CNS</i>	87.15 (13.9)	63.45 (10.6)	69.06***	0.40	–	–	–	0.54**	0.54**	0.02	–0.07	0.37**
<i>TOMM</i>	47.01 (3.1)	32.07 (8.8)	164.23***	0.62	–	–	–	–	0.34**	–0.12	0.17	–0.07
<i>WCT</i>	47.04 (3.6)	35.48 (8.9)	89.09***	0.47	–	–	–	–	–	–0.27*	–0.15	0.46**
<i>DCT</i>	11.59 (4.4)	19.16 (7.8)	38.96***	0.28	–	–	–	–	–	–	–0.31**	–0.15
<i>RDS</i>	8.51 (2.1)	6.55 (1.8)	19.42***	0.16	–	–	–	–	–	–	–	0.13
<i>RFIT</i>	12.45 (3.0)	9.59 (4.5)	14.19***	0.12	–	–	–	–	–	–	–	–

WMT Word Memory Test, *IR* immediate recognition, *DR* delayed recognition, *CNS* consistency, *TOMM* Test of Memory Malingering, *WCT* Advanced Clinical Solutions Word Choice Test, *DCT* Dot Counting Testing, *RDS* Wechsler Adult Intelligence Scale (WAIS-IV) Reliable Digit Span, *RFIT* Rey 15-Item Test. Correlational analyses total $n = 80$ for all valid participants except for *RFIT* ($n = 75$). * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 3 Failure base rates and mean Performance Validity Test failures by validity group

Performance Validity Test	Valid group failure rate	Invalid group failure rate
	<i>N</i> = 80 <i>N</i> (%)	<i>N</i> = 34 <i>N</i> (%)
WMT	4 (5)	32 (94)
TOMM T1—conservative	5 (6)	27 (79)
TOMM T1—liberal	11 (14)	30 (88)
WCT—conservative	4 (5)	23 (70)
WCT—liberal	17 (21)	26 (79)
DCT	7 (9)	23 (68)
RDS—conservative	13 (16)	15 (46)
RDS—liberal	25 (31)	21 (64)
RFIT—conservative	10 (13)	9 (29)
RFIT—liberal	21 (28)	14 (45)
	<i>M</i> (SD)	<i>M</i> (SD)
Mean PVT failures-conservative	0.53 (0.76)	3.79 (1.22)
Mean PVT failures-liberal	1.06 (1.04)	4.29 (1.36)

PVT Performance Validity Test, *WMT* Word Memory Test, *TOMM* Test of Memory Malingering, *WCT* Advanced Clinical Solutions Word Choice Test, *DCT* Dot Counting Testing, *RDS* Wechsler Adult Intelligence Scale (WAIS-IV) Reliable Digit Span, *RFIT* Rey 15-Item Test. Total *n* = 80 for all valid participants except for RFIT (*n* = 75). Total *n* = 34 for all invalid participants except WCT and RDS (*n* = 33) and RFIT (*n* = 31). Any cells with missing data are denoted in italics

combinations as well as TOMM/DCT and TOMM/WCT, whereas other $\geq 20\%$ of invalid cases passed several other PVT combinations. Overall, the WMT/TOMM, WMT/WCT, WMT/DCT, TOMM/WCT, and TOMM/DCT two-PVT combinations would yield the greatest number of participants in each validity group if used to assign group membership for a research study. Finally, the failure rate on 1/2 PVTs was high and ranged from 6 to 25% for valid cases and 26 to 68% for invalid cases. At liberal cut-scores, 53–86% of clinically valid

cases would be correctly identified as valid (i.e., failed 0/2 PVTs), with an increased false-positive rate of 0–15%, though all but 2 combinations retained a false-positive rate of $\leq 6\%$. Among invalid cases, correct classification (i.e., failure on 2/2 PVTs) significantly improved to 39–82%, and the false-negative rate ranged from 0 to 30% depending on the two-PVT combination. Again, combinations containing RDS and RFIT tended to have lower accuracy for identifying invalid cases. Similar to conservative cut-scores, the failure rate on

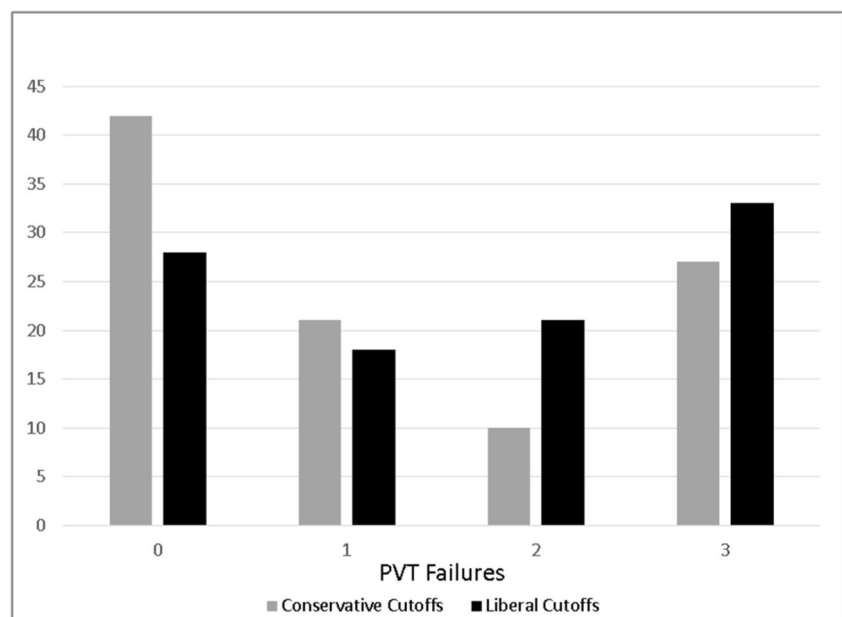
Fig. 1 Percentages of Performance Validity Test (PVT) failures using conservative versus liberal failure cut-scores

Table 4 Number of Performance Validity Test (PVT) failures by validity group for all combinations of two PVTs using conservative criterion PVT cut-scores for task failure

PVTs	Invalid (<i>n</i> =34) Number of Failures <i>n</i> (%)			Valid (<i>n</i> =80) Number of Failures <i>n</i> (%)			
	0	1	2	0	1	0-1	2
WMT/TOMM	0 (0%)	9 (26%)	25 (74%)	71 (89%)	9 (11%)	80 (100%)	0 (0%)
WMT/WCT	1 (3%)	10 (30%)	22 (67%)	72 (90%)	8 (10%)	80 (100%)	0 (0%)
WMT/DCT	0 (0%)	13 (38%)	21 (62%)	69 (86%)	11 (14%)	80 (100%)	0 (0%)
WMT/RDS	1 (3%)	18 (55%)	14 (42%)	63 (79%)	17 (21%)	80 (100%)	0 (0%)
WMT/RFIT	1 (3%)	21 (68%)	9 (29%)	61 (81%)	14 (19%)	75 (100%)	0 (0%)
TOMM/WCT	3 (9%)	11 (33%)	19 (58%)	73 (91%)	5 (6%)	78 (97%)	2 (3%)
TOMM/DCT	1 (3%)	16 (47%)	17 (50%)	68 (85%)	12 (15%)	80 (100%)	0 (0%)
TOMM/RDS	4 (12%)	17 (52%)	12 (36%)	63 (79%)	16 (20%)	79 (99%)	1 (1%)
TOMM/RFIT	4 (12%)	21 (68%)	6 (19%)	61 (82%)	13 (17%)	74 (99%)	1 (1%)
WCT/DCT	4 (12%)	13 (39%)	16 (49%)	70 (88%)	9 (11%)	79 (99%)	1 (1%)
WCT/RDS	7 (22%)	13 (41%)	12 (37%)	63 (79%)	17 (21%)	80 (100%)	0 (0%)
WCT/RFIT	8 (27%)	14 (47%)	8 (27%)	64 (85%)	8 (11%)	72 (96%)	3 (4%)
DCT/RDS	7 (21%)	14 (42%)	12 (36%)	62 (77%)	16 (20%)	68 (97%)	2 (3%)
DCT/RFIT	9 (29%)	14 (48%)	7 (23%)	59 (78%)	15 (20%)	74 (99%)	1 (1%)
RDS/RFIT	12 (40%)	13 (43%)	5 (17%)	54 (72%)	19 (25%)	73 (97%)	2 (3%)
<i>M</i> %	13.06%	45.13%	41.73%	82.73%	16.06%	98.87%	1.13%

PVT Performance Validity Test, WMT Word Memory Test, TOMM Test of Memory Malingering, WCT Advanced Clinical Solutions Word Choice Test, DCT Dot Counting Testing, RDS Wechsler Adult Intelligence Scale (WAIS-IV) Reliable Digit Span, RFIT Rey 15-Item Test. Total *n* = 80 for all valid participants except for RFIT (*n* = 75). Total *n* = 34 for all invalid participants except WCT and RDS (*n* = 33) and RFIT (*n* = 31). Any cells with missing data are denoted in italics. Shaded cells denote correct classification

1/2 PVTs remained high and ranged from 16 to 45% for valid cases and 26 to 58% for invalid cases.

At conservative cutoffs, when three PVTs were used and the benchmark for being classified as invalid was failing ≥ 2 of the 3 PVTs in the combination, classification accuracy substantially increased with 48–97% of invalid cases being correctly identified as such. (See Tables 6 and 7 for all possible three-PVT combinations using conservative and liberal cut-scores, respectively.) Moreover, the number of invalid cases which would have been incorrectly classified as valid by failing 0/3 PVTs dropped considerably and was 0 for 9/19 three-PVT combinations and ≤ 4 cases for 8/10 remaining three-PVT combinations. For valid cases, the majority (i.e., 67–86%) again failed 0 PVTs, whereas a small minority (i.e., ≤ 5 cases; 6%) failed ≥ 2 PVTs. Further examination of those valid cases who failed ≥ 2 PVTs revealed that these individuals generally had significant (e.g., Alzheimer's disease, temporal lobe epilepsy s/p resection) and/or multifactorial (e.g., multiple medical conditions with polypharmacy) cognitive impairment. Overall, the WMT/TOMM/DCT combination had the highest concordance in that it classified 33/34 (97%) invalid cases with 0/80 valid cases failing ≥ 2 of these tests, though several other three-PVT combinations also had high concordance (see Fig. 2). Finally, even when using three PVTs, 11–31% of valid cases and 3–43% of invalid cases continued to fail just one PVT. At liberal cut-scores, 60–

97% of invalid cases would be correctly identified with a false-negative rate of 0–17% depending on the three-PVT combination used. Conversely, for the valid cases, 46–75% failed 0 PVTs with a significantly increased maximum false-positive rate of 3–27%. Combinations containing RDS and RFIT again generally evidenced the weakest relative sensitivities and had the highest false-positive rates. Lastly, the failure rate on 1/2 PVTs remained high and ranged from 21 to 50% for valid cases and 3–28% for invalid cases.

Discussion

Returning to the original question of whether quality or quantity of criterion PVTs is critical for establishing validity groups in PVTs research, our results suggest that the answer is, to some degree, both. When two PVTs were used, false-negative rates of invalid cases that were incorrectly classified as valid (i.e., passed 2/2 PVTs) ranged from 0% to as high as 40% depending on the psychometric properties of the specific PVTs included in two-PVT combination. Notably, combinations of the WMT, TOMM, WCT, and DCT had the fewest false negatives, whereas RDS and RFIT resulted in higher false-negative and equivocal rates (i.e., 1 pass/1 failure) when paired with the more robust measures. This has notable implications for the choice of measures for PVT research. As a

Table 5 Number of Performance Validity Test (PVT) failures by validity group for all combinations of two PVTs using liberal criterion PVT cut-scores for task failure

PVTs	Invalid (n=34) Number of Failures n (%)			Valid (n=80) Number of Failures n (%)			
	0	1	2	0	1	0-1	2
WMT/TOMM	0 (0%)	6 (18%)	28 (82%)	66 (83%)	13 (16%)	79 (98%)	1 (1%)
WMT/WCT	1 (3%)	7 (21%)	25 (76%)	59 (74%)	21 (26%)	80 (100%)	0 (0%)
WMT/DCT	0 (0%)	13 (38%)	21 (62%)	69 (86%)	11 (14%)	80 (100%)	0 (0%)
WMT/RDS	1 (3%)	12 (36%)	20 (61%)	51 (64%)	26 (36%)	80 (100%)	0 (0%)
WMT/RFIT	0 (0%)	18 (58%)	13 (42%)	50 (67%)	25 (33%)	75 (100%)	0 (0%)
TOMM/WCT	2 (6%)	7 (21%)	24 (73%)	57 (71%)	18 (23%)	75 (94%)	5 (6%)
TOMM/DCT	1 (3%)	13 (38%)	20 (59%)	63 (79%)	16 (20%)	79 (99%)	1 (1%)
TOMM/RDS	2 (6%)	12 (36%)	19 (58%)	47 (59%)	30 (37%)	77 (96%)	3 (4%)
TOMM/RFIT	2 (7%)	17 (55%)	12 (39%)	46 (61%)	26 (35%)	72 (96%)	3 (4%)
WCT/DCT	3 (9%)	12 (36%)	18 (55%)	59 (74%)	18 (22%)	77 (96%)	3 (4%)
WCT/RDS	4 (13%)	10 (31%)	18 (56%)	41 (53%)	36 (45%)	77 (96%)	3 (4%)
WCT/RFIT	5 (17%)	12 (40%)	13 (43%)	48 (64%)	17 (23%)	65 (87%)	10 (13%)
DCT/RDS	5 (15%)	12 (36%)	16 (49%)	52 (65%)	24 (30%)	76 (95%)	4 (5%)
DCT/RFIT	9 (29%)	10 (32%)	12 (39%)	49 (65%)	24 (32%)	73 (97%)	2 (3%)
RDS/RFIT	9 (30%)	9 (30%)	12 (40%)	40 (53%)	24 (32%)	64 (85%)	11 (15%)
M %	9.40%	35.07%	55.60%	67.87%	28.27%	95.93%	4.00%

PVT Performance Validity Test, WMT Word Memory Test, TOMM Test of Memory Malingering, WCT Advanced Clinical Solutions Word Choice Test, DCT Dot Counting Testing, RDS Wechsler Adult Intelligence Scale (WAIS-IV) Reliable Digit Span, RFIT Rey 15-Item Test. Total n = 80 for all valid participants except for RFIT (n = 75). Total n = 34 for all invalid participants except WCT and RDS (n = 33) and RFIT (n = 31). Any cells with missing data are denoted in italics. Shaded cells denote correct classification

practical example, suppose a researcher conducted a hypothetical study with these data to cross-validate the WCT using the

TOMM Trial 1 and one additional PVT as the two criterion measures. As depicted in Table 8, all four receiver operating

Table 6 Number of Performance Validity Test (PVT) failures by validity group for all combinations of three PVTs using conservative criterion PVT cut-scores for task failure

Three-PVT Combination	Invalid (n=34) Number of Failures n (%)					Valid (n=80) Number of Failures n (%)				
	0	1	2	3	2+	0	1	0-1	2	3
WMT/TOMM/WCT	0 (0%)	4 (12%)	11 (33%)	18 (55%)	29 (88%)	69 (86%)	9 (11%)	78 (97%)	2 (3%)	0 (0%)
WMT/TOMM/DCT	0 (0%)	1 (3%)	18 (53%)	15 (44%)	33 (97%)	64 (80%)	16 (20%)	80 (100%)	0 (0%)	0 (0%)
WMT/TOMM/RDS	0 (0%)	5 (15%)	17 (52%)	11 (33%)	28 (85%)	59 (74%)	20 (25%)	79 (99%)	1 (1%)	0 (0%)
WMT/TOMM/RFIT	0 (0%)	5 (16%)	20 (65%)	6 (20%)	26 (85%)	57 (76%)	17 (23%)	74 (99%)	1 (1%)	0 (0%)
WMT/WCT/DCT	0 (0%)	5 (15%)	13 (39%)	14 (46%)	27 (85%)	66 (83%)	13 (16%)	79 (99%)	1 (1%)	0 (0%)
WMT/WCT/RDS	1 (3%)	6 (19%)	14 (44%)	11 (34%)	25 (78%)	59 (74%)	21 (26%)	80 (100%)	0 (0%)	0 (0%)
WMT/WCT/RFIT	0 (0%)	9 (30%)	13 (43%)	8 (27%)	21 (70%)	60 (80%)	12 (16%)	72 (96%)	3 (4%)	0 (0%)
WMT/DCT/RDS	0 (0%)	8 (24%)	14 (42%)	11 (33%)	25 (75%)	58 (73%)	20 (25%)	78 (97%)	2 (3%)	0 (0%)
WMT/DCT/RFIT	0 (0%)	10 (32%)	14 (45%)	7 (23%)	21 (68%)	55 (73%)	19 (25%)	74 (99%)	1 (1%)	0 (0%)
WMT/RDS/RFIT	0 (0%)	13 (43%)	12 (40%)	5 (17%)	17 (57%)	50 (67%)	23 (31%)	73 (97%)	2 (3%)	0 (0%)
TOMM/WCT/DCT	1 (3%)	5 (15%)	15 (46%)	12 (36%)	27 (82%)	67 (84%)	10 (13%)	77 (96%)	3 (4%)	0 (0%)
TOMM/WCT/RDS	2 (6%)	8 (25%)	12 (38%)	10 (31%)	22 (69%)	61 (76%)	16 (20%)	77 (96%)	3 (4%)	0 (0%)
TOMM/WCT/RFIT	2 (7%)	9 (30%)	13 (43%)	6 (20%)	19 (63%)	61 (81%)	10 (13%)	71 (95%)	3 (4%)	1 (1%)
TOMM/DCT/RDS	1 (3%)	9 (27%)	14 (42%)	9 (27%)	23 (69%)	58 (72%)	19 (24%)	77 (96%)	3 (4%)	0 (0%)
TOMM/DCT/RFIT	1 (3%)	11 (36%)	15 (48%)	4 (13%)	19 (61%)	55 (73%)	18 (24%)	73 (97%)	2 (3%)	0 (0%)
WCT/DCT/RDS	4 (13%)	6 (19%)	13 (41%)	9 (28%)	22 (69%)	59 (74%)	18 (22%)	77 (96%)	4 (4%)	0 (0%)
WCT/DCT/RFIT	4 (13%)	9 (30%)	11 (37%)	6 (20%)	17 (57%)	58 (77%)	14 (19%)	72 (96%)	2 (3%)	1 (1%)
WCT/RDS/RFIT	5 (17%)	10 (35%)	9 (31%)	5 (17%)	14 (48%)	53 (71%)	17 (23%)	70 (94%)	5 (6%)	0 (0%)
DCT/RDS/RFIT	7 (23%)	6 (20%)	14 (47%)	3 (10%)	17 (57%)	50 (67%)	20 (27%)	70 (94%)	5 (6%)	0 (0%)
M %	4.79%	23.47%	43.63%	28.11%	71.73%	75.84%	21.21%	97.05%	2.89%	0.11%

PVT Performance Validity Test, WMT Word Memory Test, TOMM test of Memory Malingering, WCT Advanced Clinical Solutions Word Choice Test, DCT Dot Counting Testing, RDS Wechsler Adult Intelligence Scale (WAIS-IV) Reliable Digit Span, RFIT Rey 15-Item Test. Total n = 80 for all valid participants except for RFIT (n = 75). Total n = 34 for all invalid participants except WCT and RDS (n = 33) and RFIT (n = 31). Any cells with missing data are denoted in italics. Shaded cells denote correct classification

Table 7 Number of Performance Validity Test (PVT) failures by validity group for all combinations of three PVTs using liberal criterion PVT cut-scores for task failure

Three-PVT Combination	Invalid (n=34) Number of Failures n (%)					Valid (n=80) Number of Failures n (%)				
	0	1	2	3	2+	0	1	0-1	2	3
WMT/TOMM/WCT	0 (0%)	3 (9%)	7 (21%)	23 (70%)	30 (91%)	54 (68%)	20 (35%)	74 (92%)	6 (8%)	0 (0%)
WMT/TOMM/DCT	0 (0%)	1 (3%)	15 (44%)	18 (53%)	33 (97%)	60 (75%)	18 (22%)	78 (97%)	2 (3%)	0 (0%)
WMT/TOMM/RDS	0 (0%)	3 (9%)	12 (36%)	18 (55%)	30 (91%)	44 (55%)	32 (40%)	76 (95%)	4 (5%)	0 (0%)
WMT/TOMM/RFIT	0 (0%)	2 (7%)	18 (58%)	11 (35%)	29 (93%)	43 (57%)	28 (37%)	71 (99%)	4 (5%)	0 (0%)
WMT/WCT/DCT	0 (0%)	4 (12%)	12 (36%)	17 (52%)	29 (88%)	55 (69%)	22 (27%)	77 (96%)	3 (4%)	0 (0%)
WMT/WCT/RDS	1 (3%)	3 (9%)	11 (34%)	17 (53%)	28 (87%)	37 (46%)	40 (50%)	77 (96%)	3 (4%)	0 (0%)
WMT/WCT/RFIT	0 (0%)	5 (17%)	13 (43%)	12 (40%)	25 (77%)	44 (59%)	21 (28%)	65 (87%)	10 (13%)	0 (0%)
WMT/DCT/RDS	0 (0%)	6 (18%)	12 (36%)	15 (46%)	27 (82%)	48 (60%)	28 (35%)	76 (95%)	4 (5%)	0 (0%)
WMT/DCT/RFIT	0 (0%)	9 (28%)	11 (36%)	11 (36%)	22 (72%)	45 (60%)	28 (37%)	73 (97%)	2 (3%)	0 (0%)
WMT/RDS/RFIT	0 (0%)	9 (30%)	10 (33%)	11 (37%)	21 (70%)	36 (48%)	28 (37%)	64 (85%)	11 (15%)	0 (0%)
TOMM/WCT/DCT	1 (3%)	3 (9%)	13 (39%)	16 (49%)	29 (88%)	54 (68%)	17 (21%)	71 (89%)	9 (11%)	0 (0%)
TOMM/WCT/RDS	1 (3%)	5 (16%)	9 (28%)	17 (53%)	26 (81%)	37 (46%)	34 (43%)	71 (89%)	8 (10%)	1 (1%)
TOMM/WCT/RFIT	1 (3%)	6 (20%)	11 (37%)	12 (40%)	23 (77%)	43 (57%)	18 (24%)	61 (81%)	12 (16%)	2 (3%)
TOMM/DCT/RDS	1 (3%)	5 (15%)	13 (39%)	14 (42%)	27 (81%)	44 (55%)	30 (38%)	74 (93%)	5 (6%)	1 (1%)
TOMM/DCT/RFIT	1 (3%)	2 (29%)	11 (36%)	10 (32%)	21 (68%)	42 (56%)	27 (36%)	69 (92%)	6 (8%)	0 (0%)
WCT/DCT/RDS	2 (6%)	6 (19%)	10 (31%)	14 (44%)	24 (75%)	40 (50%)	32 (40%)	72 (90%)	7 (9%)	1 (1%)
WCT/DCT/RFIT	3 (10%)	8 (27%)	8 (27%)	11 (36%)	19 (63%)	44 (59%)	20 (27%)	64 (85%)	9 (12%)	2 (3%)
WCT/RDS/RFIT	3 (10%)	7 (24%)	8 (28%)	11 (38%)	19 (66%)	35 (47%)	20 (26%)	55 (73%)	18 (24%)	2 (3%)
DCT/RDS/RFIT	5 (17%)	7 (23%)	8 (27%)	10 (33%)	18 (60%)	38 (51%)	22 (29%)	60 (80%)	14 (19%)	1 (1%)
M %	3.21%	17.05%	35.21%	44.42%	79.63%	57.16%	33.26%	90.42%	9.47%	0.68%

PVT Performance Validity Test, WMT Word Memory Test, TOMM Test of Memory Malingering, WCT Advanced Clinical Solutions Word Choice Test, DCT Dot Counting Testing, RDS Wechsler Adult Intelligence Scale (WAIS-IV) Reliable Digit Span, RFIT Rey 15-Item Test. Total n = 80 for all valid participants except for RFIT (n = 75). Total n = 34 for all invalid participants except WCT and RDS (n = 33) and RFIT (n = 31). Any cells with missing data are denoted in italics. Shaded cells denote correct classification

characteristic (ROC) curve analyses would yield significant areas under the curve (AUCs); however, alternating the two criterion PVTs would result in (1) different optimal cut-scores, (2) a 23% difference in sensitivity if ≥ 90% specificity is to be maintained (per recommendation by Boone (2012) and Larrabee et al. (2019)), and (3) a large range in the number of cases with one only criterion PVT failure (i.e., 18–33%) who may be excluded from

analysis depending on the specific two-PVT criterion combination. Thus, both quantity and quality of measures used as criterion measures in PVT research can greatly affect the reliability and replicability of study findings, and, by extension, evidenced-base clinical practice. These methodological decisions may also limit the use of historical datasets if appropriate PVTs were not included at the time of data collection.

Fig. 2 Classification accuracy of three-Performance Validity Test (PVT) combinations for identifying valid and invalid cases using conservative scores

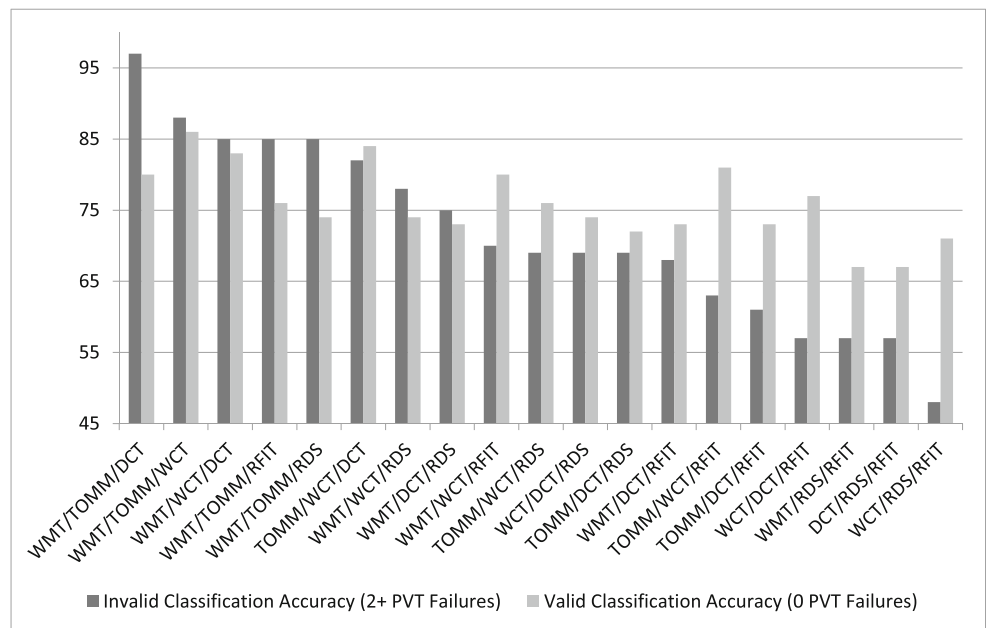


Table 8 Hypothetical study cross-validating the word choice test using different Performance Validity Test combinations to establish validity groups

PVT combination	Valid <i>N</i>	Invalid <i>N</i>	AUC	Cutoff	Sensitivity/ specificity	1 PVT failure-excluded <i>N</i>
TOMM/WMT (<i>N</i> = 113)	71	24	0.95***	≤ 43	0.83/0.91	18
TOMM/DCT (<i>N</i> = 113)	69	16	0.94***	≤ 44	0.81/.90	28
TOMM/RDS (<i>N</i> = 112)	67	13	0.92***	≤ 42	0.77/0.92	32
TOMM/RFIT (<i>N</i> = 105)	65	7	0.99***	≤ 43	1.0/0.92	33

PVT Performance Validity Test, TOMM Test of Memory Malingering, WMT Word Memory Test, DCT Dot Counting Testing, RDS Wechsler Adult Intelligence Scale (WAIS-IV) Reliable Digit Span, RFIT Rey 15-Item Test, AUC area under the curve. ****p* < .001

When the number of PVTs was increased from two to three, overall false-negative rates dropped considerably, such that 13/19 possible three-PVT combinations had ≤ 3% of invalid cases failing 0/3 PVTs. WMT/TOMM/DCT classified the most participants given all but one invalid case failed ≥ 2/3 of these tests and 0/80 valid cases failed ≥ 2/3. While increasing the number of PVTs from two to three reduced false negatives overall, it is important to note that 4/19 of three-PVT combinations still had misclassification rates of > 10%, with the combination on DCT/RDS/RFIT having this highest false-negative rate (i.e., 23%) followed by WCT/RDS/RFIT (17%), indicating PVT quality still matters, to some degree, even when more PVTs are used.

Minimizing valid participants misclassified as invalid is equally paramount in PVT research. To this end, when two PVTs were used, the false-positive rate (i.e., failed 2/2) was low (i.e., 0–4%). When the number of PVTs increased to three, the false-positive rate remained low (i.e., 0–6%), which is consistent with prior findings that document a small, but nonsignificant increase in the false-positive rate as the number of PVTs increase (Larrabee, 2014; Critchfield et al., 2019). As specificity of these PVT cut-scores (or any combination derived from individual PVT cut-scores for that matter) is less than 100% if adequate sensitivity is to be maintained, there is a risk that a small percentage with cognitive impairment will be misidentified as invalid. PVT combinations also could yield different classification accuracy in clinical populations with different levels of impairment severity. Specifically, combinations including RDS and RFIT in elderly patients and patients with suspected/known cognitive impairment are cautioned, and clinicians are encouraged to remain abreast to the literature focusing on alternative scoring/cutoff criteria if these measures are used (Bailey et al., 2018a; Fazio et al., 2019; Poynter et al., 2019; Webber et al., 2018b; Webber & Soble, 2018). Consequently, using appropriate, empirically derived PVT cut-scores within the context of clinical history and structured validity criteria may help reduce the likelihood that these individuals are included in noncredible reference samples when conducting PVT research.

This study also emphasized the research methodological challenges when a participant has just one PVT failure.

Notably, regardless of whether two- or three-PVT combinations were used, several cases in both validity groups failed only one PVT. Given all participants in this study had at least seven PVTs during their evaluation along with corroborating clinical data, clinical classification of overall validity status was relatively straightforward. In contrast, validity classification becomes more complex in a research context if all the investigator has available to classify validity status is a two- or three-PVT combination, particularly given concern that a sizeable percentage of participants would end up in the wrong reference group if they were retained via a uniform operational definition (e.g., those with one failure were classified as invalid). While a simple solution is excluding these participants from research studies, doing so may reduce generalizability of results to mixed clinical samples, particularly considering that many of the valid cases with one PVT failure in our sample had diagnosed cognitive impairment. Thus, depending on the population of interest, using alternative operational criteria that examine PVT scores in the context of other available clinical information and/or established criteria for noncredible performance (e.g., Slick criteria; Slick et al., 1999) to more accurately establish group assignment may be advisable.

Finally, given recent research (e.g., An et al., 2017) has suggested potential utility of applying liberal PVT cutoffs, this study examined the effect of adopting more liberal TOMM, WCT, RDS, and RFIT cut-scores on overall validity classification rates. For two-PVT combinations, accuracy for identifying invalid cases increased from 17–74% at conservative cutoffs to 39–85% at liberal cutoffs, whereas the decrease in false negatives was less pronounced for liberal (i.e., 0–30%) versus conservative (i.e., 0–40%) cutoffs. By contrast, liberal cutoffs resulted in significantly more valid cases misclassified as invalid (i.e., 0–15% versus 0–3%) as well as a notable drop (i.e., 72–91% to 53–86%) in the number of valid cases with 0/2 PVTs relative to conservative cutoffs. Similarly, when liberal cutoffs were used for three-PVT combinations, overall accuracy for identifying invalid cases increased minimally from 57 to 97% at conservative cut-scores to 60–97%, with a corresponding false negative decrease from 0–23% to 0–17%. Moreover, among valid cases, the use of liberal cutoffs significantly increased the false-positive rate from 0–6% to 3–

27%, while simultaneously reducing the number of valid cases with 0 PVT failures from 67–86% to 46–75%. Thus, a modest increase in sensitivity associated with liberal cutoffs was countered with a significant reduction in specificity with nearly half of three-PVT combinations having a $\geq 10\%$ false-positive rate. Across both two- and three-PVT combinations, those containing RDS and RFIT generally continued to evidence the weakest relative classification accuracy even after adopting liberal cutoffs, and, particularly in the case of three-PVT combinations, produced substantially higher false-positive rates that exceeded the 90% specificity benchmark (Boone, 2012). Consequently, results indicated that PVT quality remains critically important for overall classification, even in the context of adjusting cutoffs more liberally, or specificity will be sacrificed for, at best, modest gains in sensitivity. Finally, even with liberal scores, 18–55% of invalid and 14–45% of valid cases failed one PVT for two-PVT combinations, and 3–28% of invalid cases and 21–43% of valid cases failed just one PVT for three-PVT combinations. Therefore, despite applying more liberal cutoffs, a large number of valid and invalid cases continued to fail only one PVT in each combination, which is especially problematic for two-PVT combinations as this would result in a considerable number of essentially equivocal cases with one PVT pass and one PVT failure if only two PVTs were available as criterion measures.¹

Whether a researcher is attempting to test the classification accuracy of a PVT or rule out the impact of validity status on other neuropsychological tests (e.g., Webber et al., 2018c), current results are particularly relevant for research studies that incorporate information on performance validity status. While clinicians may have the liberty of adding additional PVTs *in vivo* to clearly establish a patient's validity status during clinical and/or forensic exams, research studies in which test selection was established *a priori* often will have more limited flexibility. When validating new PVTs, results suggest a three-PVT

¹ One reviewer remarked that a single PVT failure using liberal cutoffs is not equivocal, but rather should be considered valid. While this position is tenable in a context in which multiple PVTs are administered, we maintain that when only two PVTs are given, one failure is, by definition, equivocal in that other extra-test data would ultimately need to be considered to establish validity status. Our objective data clearly demonstrated that even when liberal cutoffs were applied, for two-PVT combinations, mean failure rates of 1/2 PVTs were 35% for invalid cases and 28% for valid cases (Tables 4 and 5). Thus, nearly 30% of invalid cases would be misclassified if cases with one failure were automatically classified as valid. Additionally, in response to the significant increase in false positives for the three-PVT combinations using liberal cutoffs (i.e., $> 10\%$ on 7/19 combinations), the reviewer suggested that the solution to improve specificity (i.e., $\geq 90\%$; Boone, 2012) was to raise the invalidity threshold to 3/3 PVT failures. However, this approach is not consistent with current practice standards, in which ≥ 2 failures is the generally accepted benchmark for identifying probable invalidity (Larrabee, 2014; Meyers & Volbrecht, 2003), and would result in an unacceptable decrease in overall mean sensitivity from 79 to 35% for identifying invalid cases, whereas using conservative cutoffs and retaining the well-established ≥ 2 failures benchmark yielded 72% mean sensitivity while maintaining 97% mean specificity and 0/19 combinations with a false-positive rate above 6% (see Tables 6 and 7).

combination efficiently maximizes classification accuracy over a two-PVT combination. Using only two PVTs is likely to decrease overall classification accuracy and introduce meaningful error when only one of the two measures are failed. For instance, a large number of participants failed one PVT in the combination regardless of whether a two- or three-PVT combination was used. With a two-PVT combination, a researcher's selection would rely upon a single score, as either the passing score or failing score was ultimately weighed as most important to group assignment, which has been recently identified as having significant limitations (see Lippa, 2018 and Schroeder et al., 2019). In contrast, a three-PVT combination likely would provide additional clarity to establish validity groups as this would allow the research to use ≥ 2 independent PVT failures as the benchmark for invalid performance, which mirrors recommended clinical practice guidelines (Larrabee, 2014).

Of note, the strength of this study is the inclusion of a diverse, mixed clinical sample, which served to enhance the generalizability of findings. In contrast to other studies which have reported increased likelihood of false positives in subjects whose cultural identity may differ from the majority culture (e.g., Gasquoin et al., 2017), no such differences in classification were identified in the current study. Further, consideration of alternative cutoffs validated in elderly patients, or those with cognitive impairment (e.g., WMT GMIP), reduced false-positive errors, and underscores the importance of considering these characteristics in research designs.

Limitations and Future Directions

While this study makes meaningful progress toward identifying the best methods for establishing criterion groups in PVT research, future research is still needed to identify the most appropriate test combinations and whether varying the type of PVTs provides the optimal classification accuracy among different clinical samples. While our clinical sample supports use of memory, attention, and speed-based PVTs as the foundation for a robust PVT combination, it is a limitation that the majority of the PVTs in this study were memory-based (or appeared to assess memory). Replication of our results with non-memory-based PVTs is warranted considering that some patients may selectively exhibit invalid performance for a cognitive domain (Boone, 2009, 2012). Therefore, the optimal combination of criterion PVTs may not only include PVTs that measure (or appear to measure) multiple cognitive domains, but also include construct match between the measure being validated and the criterion PVTs.

Given different published cut-scores exist for many PVTs, individual study findings may be, to some degree, dependent on and/or may change as a function of the specific pass/failure cut-scores used for each PVT. Unsurprisingly, prior research has yielded varying published cut-scores for several PVTs in

the current study. As such, this study offers an important reminder that interpretation of PVTs (as well as the PVTs efficacy or “quality”) depends on the pass/fail cut-scores selected; the choice of which is ultimately at the professional’s discretion. This is a particular issue with newer PVTs (e.g., WCT), in which there are comparatively fewer studies and a lack of systematic reviews/meta-analyses available to guide optimal cut-score selection. Nonetheless, data from this study highlighted how simply using more liberal cutoffs may not necessarily be the optimal solution to this problem given that more liberal cutoffs predictably increased accurate identification of invalid cases, but came with a cost of having significantly higher false positives, while continuing to maintain a substantial number of individuals with one PVT failure among both valid and invalid cases. Another limitation is that the PVTs examined in this study comprised the majority of those that were originally used to establish validity groups during the clinical evaluation. Therefore, cross-validation of these results in other clinical samples and/or utilizing a prospective study methodology with independent criterion PVTs for establishing validity groups that are separate/distinct from the test PVT combinations being examined will be critical in replication efforts.

Although our overall false-positive rate was low and consistent with prior research that similarly found a low false-positive rate in Alzheimer’s disease when failure on 2 PVTs was used as the criterion (Loring et al., 2016), our mixed clinical sample included cognitively impaired patients with predominantly mild presentations. Replication of findings in clinical samples with a higher prevalence of more severe cognitive impairment/dementia is recommended to increase external validity. Relatedly, the use of a veteran sample may limit the generalizability of results to other populations given unique aspects of this population (Armistead-Jehle et al., 2017). Unlike civilian populations, external incentive (i.e., service-connection disability) is an ever present concern when delivering neuropsychological services in the VA system, particularly considering that evaluation for VA service connected-disability initiation or rating increase can be requested at any point and past research demonstrating that PVT performance can vary with evaluation context in veteran and active duty populations (e.g., clinical vs. disability, history of military concussion; Armistead-Jehle et al., 2016; Grills & Armistead-Jehle, 2016). Finally, it is a limitation that no more than three-PVT combinations were examined as increased accuracy may be obtained with more measures.

While it is well-established that failing ≥ 2 PVTs accurately identifies those exhibiting noncredible performance (Larrabee, 2008; Meyers et al., 2014), few studies have examined how using multiple PVTs to identify criterion groups in PVT research impacts study findings. This study builds on the work of Schroeder et al. (2019) and demonstrates that both the number and the quality of PVTs are important when

operationalizing criterion variables in PVT research. Base rate of PVT failure and clinical history can be useful for identifying invalidity clinically, though researchers may be tasked with selecting PVTs to serve as criterion measures a priori. By providing evidence that both the number and quality of PVTs impact criterion grouping, these results may help inform selection of criterion PVTs and have downstream effects on increased replicability of PVT research findings.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Disclaimer The views expressed herein are those of the authors and do not necessarily reflect the views or the official policy of the Department of Veterans Affairs or US Government.

References

- Alverson, W. A., O’Rourke, J. J. F., & Soble, J. R. (2019). The word memory test genuine memory impairment profile discriminates genuine memory impairment from invalid performance in a mixed clinical sample with cognitive impairment. *The Clinical Neuropsychologist*.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders - fifth edition (DSM-5)*. Washington, DC: American Psychiatric Publishing.
- An, K. Y., Kaploun, K., Erdodi, L. A., & Abeare, C. A. (2017). Performance validity in undergraduate research participants: a comparison of failure rates across tests and cutoffs. *The Clinical Neuropsychologist*, 31, 193–206.
- Armistead-Jehle, P., Cooper, D. B., & Vanderploeg, R. D. (2016). The role of performance validity tests in the assessment of cognitive functioning after military concussion: a replication and extension. *Applied Neuropsychology: Adult*, 23(4), 264–273.
- Armistead-Jehle, P., Soble, J. R., Cooper, D. C., & Belanger, H. G. (2017). Unique aspects of TBI in military populations. [Special issue]. *Physical Medicine & Rehabilitation Clinics of North America*, 28, 323–337.
- Bailey, K. C., Soble, J. R., Bain, K. M., & Fullen, C. (2018a). Embedded performance validity tests in the Hopkins verbal learning test – revised and the brief visuospatial memory test – revised: a replication study. *Archives of Clinical Neuropsychology*, 33, 895–900.
- Bailey, K. C., Soble, J. R., & O’Rourke, J. J. F. (2018b). Clinical utility of the Rey 15-Item Test, recognition trial, and error scores for detecting noncredible neuropsychological performance validity in a mixed-clinical sample of veterans. *The Clinical Neuropsychologist*, 32, 119–131.
- Bain, K. M., & Soble, J. R. (2019). Validation of the advanced clinical solutions word choice test (WCT) in a mixed clinical sample: establishing classification accuracy, sensitivity/specificity, and cutoff scores. *Assessment*, 26, 1320–1328.
- Bain, K. M., Soble, J. R., Webber, T. A., Messerly, J. M., Bailey, K. C., Kirton, J. W., & McCoy, K. J. M. (2019). Cross-validation of three advanced clinical solutions performance validity tests: examining combinations of measures to maximize classification of invalid performance. *Applied Neuropsychology: Adult*.
- Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist*, 23, 729–741.

- Boone, K. B. (2012). *Clinical practice of forensic neuropsychology*. New York: Guilford Press.
- Boone, K., Lu, P., & Herzberg, D. S. (2002). *The dot counting test manual*. Los Angeles: Western Psychological Services.
- Critchfield, E. A., Soble, J. R., Marceaux, J. C., Bain, K. M., Bailey, K. C., Webber, T. A., et al. (2019). Cognitive impairment does not cause performance validity failure: analyzing performance patterns among unimpaired, impaired, and noncredible participants across six tests. *The Clinical Neuropsychologist*, 6, 1083–1101.
- Denning, J. H. (2012). The efficiency and accuracy of the Test of Memory Malingering trial 1, errors on the first 10 items of the Test of Memory malingering, and five embedded measures in predicting invalid test performance. *Archives of Clinical Neuropsychology*, 27(4), 417–432.
- Erdodi, L. A. (2019). Aggregating validity indicators: the salience of domain specificity and the indeterminate range in multivariate models of performance validity assessment. *Applied Neuropsychology: Adult*, 26(2), 155–172.
- Erdodi, L. A., Kirsch, N. L., Lajiness-O'Neill, R., Vingilis, E., & Medoff, B. (2014). Comparing the Recognition Memory Test and the Word Choice Test in a mixed clinical sample: are they equivalent? *Psychological Injury and Law*, 7, 255–263.
- Fazio, R. L., Farris, A. N., & Yamout, K. Z. (2019). Use of the Rey 15-Item Test as a performance validity test in an elderly population. *Applied Neuropsychology: Adult*, 26, 28–35.
- Gasquoin, P. G., Weimer, A. A., & Amador, A. (2017). Specificity rates for non-clinical, bilingual, Mexican Americans on three popular performance validity measures. *The Clinical Neuropsychologist*, 31(3), 587–597. <https://doi.org/10.1080/13854046.2016.1277786>.
- Green, P. (2003). *Green's word memory test for windows: user's manual*. Edmonton: Green's Publishing.
- Green, P., Montijo, J., & Brockhaus, R. (2011). High specificity of the Word Memory Test and Medical Symptom Validity Test in groups with severe verbal memory impairment. *Applied Neuropsychology*, 18(2), 86–94.
- Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingering amnesia measures with a large clinical sample. *Psychological Assessment*, 6(3), 218–224.
- Grills, C. E., & Armistead-Jehle, P. (2016). Performance validity test and neuropsychological assessment battery screening module performances in an active-duty sample with a history of concussion. *Applied Neuropsychology: Adult*, 23(4), 295–301.
- Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: relationship to likelihood ratios. *The Clinical Neuropsychologist*, 22(4), 666–679.
- Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology*, 29(4), 364–373.
- Larrabee, G. J., Rohling, M. L., & Meyers, J. E. (2019). Use of multiple performance and symptom validity measures: determining the optimal per test cutoff for determination of invalidity, analysis of skew, and inter-test correlations in valid and invalid performance groups. *The Clinical Neuropsychologist*, 1–19.
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). Oxford: Oxford University Press.
- Lippa, S. M. (2018). Performance validity testing in neuropsychology: a clinical guide, critical review, and update on a rapidly evolving literature. *The Clinical Neuropsychologist*, 32(3), 391–421.
- Loring, D. W., Goldstein, F. C., Chen, C., Drane, D. L., Lah, J. J., Zhao, L., & Larrabee, G. J. (2016). False-positive error rates for Reliable Digit Span and Auditory Verbal Learning Test performance validity measures in amnesic mild cognitive impairment and early Alzheimer disease. *Archives of Clinical Neuropsychology*, 31(4), 313–331.
- Martin, P. K., Schroeder, R. W., Olsen, D. H., Maloy, H., Boettcher, A., Ernst, N., & Okut, H. (2019). A systematic review and meta-analysis of the Test of Memory Malingering in adults: two decades of deception detection. *The Clinical Neuropsychologist*. <https://doi.org/10.1080/13854046.2019.1637027>.
- Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: a survey of North American professionals. *The Clinical Neuropsychologist*, 29(6), 741–776.
- Meyers, J. E., Miller, R. M., Thompson, L. M., Scalse, A. M., Allred, B. C., Rupp, Z. W., Dupaix, Z. P., & Junghyun Lee, A. (2014). Using likelihood ratios to detect invalid performance with performance validity measures. *Archives of Clinical Neuropsychology*, 29(3), 224–235.
- Meyers, J. E., & Volbrecht, M. (2003). A validation of multiple malingering detection methods in a large clinical sample. *Archives of Clinical Neuropsychology*, 18, 261–276.
- Novitski, J., Steele, S., Karantzoulis, S., & Randolph, C. (2012). The repeatable battery for the assessment of neuropsychological status effort scale. *Archives of Clinical Neuropsychology*, 27, 190–195.
- Pearson. (2009). *Advanced clinical solutions for WAIS-IV and WMS-IV: clinical and interpretive manual*. San Antonio: Pearson.
- Poreh, A., Bezdicek, O., Korobkova, I., Levin, J. B., & Dines, P. (2016). The Rey Auditory Verbal Learning Test forced-choice recognition task: base-rate data and norms. *Applied Neuropsychology: Adult*, 23, 155–161.
- Poynter, K., Boone, K. B., Ermshar, A., Miora, D., Cottingham, M., Victor, T. L., Ziegler, E., Zeller, M. A., & Wright, M. (2019). Wait, there's a baby in this bath water! Update on quantitative and qualitative cut-offs for Rey 15-Item Recall and Recognition. *Archives of Clinical Neuropsychology*, 34, 1367–1380.
- Rai, J. K., & Erdodi, L. A. (2019). Impact of criterion measures on the classification accuracy of TOMM-1. *Applied Neuropsychology: Adult*.
- Rey, A. (1964). *L'examen Clinique en psychologie*. Paris: Presses Universitaires de France.
- Schroeder, R. W., Martin, P. K., Heinrichs, R. J., & Baade, L. E. (2019). Research methods in performance validity testing studies: criterion grouping approach impacts study outcomes. *The Clinical Neuropsychologist*, 33, 466–477.
- Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable digit span: a systematic review and cross-validation study. *Assessment*, 19(1), 21–30.
- Schwartz, E. S., Erdodi, L., Rodriguez, N., Ghosh, J. J., Curtain, J. R., Flashman, L. A., & Roth, R. M. (2016). CVLT-II forced choice recognition trial as an embedded validity indicator: a systematic review of the evidence. *Journal of the International Neuropsychological Society*, 22, 851–858.
- Silverberg, N. D., Wertheimer, J. C., & Fichtenberg, N. L. (2007). An effort index for the Repeatable Battery For The Assessment Of Neuropsychological Status (RBANS). *The Clinical Neuropsychologist*, 21(5), 841–854.
- Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingering neurocognitive dysfunction: proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, 13(4), 545–561.
- Slick, D. J., Tan, J. E., Strauss, E. H., & Hultsch, D. F. (2004). Detecting malingering: a survey of experts' practices. *Archives of Clinical Neuropsychology*, 19(4), 465–473.
- Tombaugh, T. N. (1996). *Test of memory malingering (TOMM)*. North Tonawanda: Multi Health Systems.
- Webber, T. A., Bailey, K. C., Alverson, W. A., Critchfield, E. A., Bain, K. M., Messerly, J. M., et al. (2018a). Further validation of the Test of Memory Malingering (TOMM) Trial 1: examination of false positives and convergence with other validity measures. *Psychological Injury and Law*, 11, 325–335.

- Webber, T. A., Critchfield, E. A., & Soble, J. R. (2018b). Convergent, discriminant, and concurrent validity of non-memory-based performance validity tests. *Assessment*.
- Webber, T. A., Marceaux, J. C., Critchfield, E. A., & Soble, J. R. (2018c). Relative impacts of mild and major neurocognitive disorder on rate of verbal learning acquisition. *Archives of Clinical Neuropsychology*.
- Webber, T. A., & Soble, J. R. (2018). Utility of various WAIS-IV digit span indices for identifying noncredible performance validity among cognitively impaired and unimpaired examinees. *The Clinical Neuropsychologist*, 32(4), 657–670.
- Wechsler, D. (2008). *WAIS-IV: administration and scoring manual*. San Antonio: Pearson.
- Weinstein, S., Obuchowski, N. A., & Lieber, M. L. (2005). Clinical evaluation of diagnostic tests. *American Journal of Roentgenology*, 184(1), 141–149.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.