Published in final edited form as:

J Clin Exp Neuropsychol. 2018 December ; 40(10): 1013–1021. doi:10.1080/13803395.2018.1472221.

Performance Validity in Older Adults: Observed Versus Predicted False Positive Rates in Relation to Number of Tests Administered

Jeremy J. Davis

University of Utah School of Medicine, Division of Physical Medicine & Rehabilitation, 30 North 1900 East, Salt Lake City, UT 84132, Phone: 801-581-2932

Abstract

Introduction—This study examined false positive rates on embedded performance validity tests (PVTs) in older adults grouped by cognitive status.

Method—The research design involved secondary analysis of data from the National Alzheimer's Coordinating Center database. Participants ($N = 22,688$) were grouped by cognitive status: normal ($n = 10,319$), impaired ($n = 1,194$), amnesic or non-amnesic mild cognitive impairment (MCI; $n = 5,414$), and dementia ($n = 5,761$). Neuropsychological data were used to derive 5 PVTs.

Results—False positive rates on individual PVTs ranged from 3.3 to 26.3% with several embedded PVTs showing acceptable specificity across groups. The proportion of participants failing two or more PVTs varied by cognitive status: normal (1.9%), impaired (6.6%), MCI (13.2%), and dementia (52.8%). Comparison of observed and predicted false positive rates at different specificity levels (.85 or .90) demonstrated significant differences in all comparisons. In normal and impaired groups, predicted rates were higher than observed rates. In the MCI group, predicted and observed comparisons varied: predicted rates were higher with specificity at .85 and lower with specificity at .90. In the dementia group, predicted rates underestimated observed rates.

Conclusions—Despite elevated false positives in conditions involving severe cognitive compromise, several measures retain acceptable specificity regardless of cognitive status. Predicted false positive rates based on the number of PVTs administered were not observed empirically. These findings do not support the utility of simulated data in predicting false positive rates in older adults.

Keywords

Performance validity; false positive rate; older adults; dementia

Performance validity tests (PVTs) are commonly used by neuropsychologists in clinical and forensic contexts to estimate the likelihood that obtained test data accurately reflect ability (Bush et al., 2005; Heilbronner et al., 2009; Sharland & Gfeller, 2006). There has been considerable research activity in the area over the last two decades resulting in a proliferation of tests covering almost all cognitive domains (Boone, 2009; Sweet & Guidotti Breting, 2013). Research has also examined underlying constructs, models, and nomenclature (Larrabee, 2012; Merkelbach, Jelicic, & Pieters, 2011; Van Dyke, Millis, Axelrod, & Hanks, 2013). Given the weight attributed to information provided by PVTs and

their frequent use in medicolegal contexts, clinical and research questions on performance validity often spark controversy (Faust & Guilmette, 1990).

Among other areas of contention in the PVT literature, issues related to PVT failure and false positive rates are a recurrent concern (Berthelson, Mulchan, Odland, Miller, & Mittenberg, 2013; Davis & Millis, 2014; Larrabee, 2014). A false positive on a PVT occurs when an individual is identified by the test as performing noncredibly or invalidly (i.e., with what was previously described as inadequate effort; Larrabee, 2012) when, in fact, that individual was providing a credible or valid (i.e., best effort) response. A number of demographic variables have been identified as relevant to PVT failure including age (Duff et al., 2011), educational history (Webb, Batchelor, Meares, Taylor, & Marsh, 2012), and ethnicity (Salazar, Lu, Wen, & Boone, 2007). Clinical conditions have also been identified as important considerations regarding PVT failure including intellectual disability (Dean, Victor, Boone, & Arnold, 2008) and dementia (Dean, Victor, Boone, Philpott, & Hess, 2009).

The recognition memory testing format used by many freestanding PVTs can be sensitive to memory deficits in dementia. Among the older freestanding PVTs in common use today, the Test of Memory Malingering (TOMM; Tombaugh, 1996) illustrates this issue. In healthy, nonclinical volunteers, TOMM performances are often nearly perfect across the last two trials (Trial 2: $M = 49.9$, $SD = 0.4$; Retention: $M = 49.9$, $SD = 0.5$; Tombaugh, 1996), but in individuals with a diagnosis of dementia and no identified incentive to underperform, performances drop below cutoff (Trial 2: $M = 39.5$, $SD = 6.1$; Retention: $M = 39.5$, $SD = 6.8$; Teichner & Wagner, 2004) and demonstrate unacceptably low specificity (.50 to .70; Merten, Bossink, & Schmand, 2007). Similarly elevated false positive rates have been reported on other freestanding PVTs (Merten et al., 2007), but consideration of performance patterns from easier to more demanding trials has led to better classification even in memory disorders clinics (Howe & Loring, 2009).

Embedded PVTs have also been found to have unacceptably high false positive rates in cases involving dementia (Dean et al., 2009). Dean and colleagues examined 18 embedded and freestanding PVTs derived from 12 neuropsychological measures in a clinical sample of participants diagnosed with dementia. Only two embedded PVTs based on Digit Span (Wechsler, 1987b) showed adequate specificity. Loring et al. (2016) examined embedded PVTs on Digit Span and the Auditory Verbal Learning Test (AVLT; Rey, 1941; Ivnik et al., 1992) in a sample of older adults categorized by cognitive status into intact, mild cognitive impairment (MCI), and dementia groups. A commonly used reliable digit span (Greiffenstein, Baker, & Gola, 1994) cutoff showed adequate specificity in the cognitively intact group but the false positive rate increased to 14% and 34% in MCI and dementia groups, respectively. On the AVLT, a recognition trial cutoff showed adequate specificity in the cognitively intact group, but the false positive rate was 42% and 70% in MCI and dementia groups, respectively. A logistic regression model cutoff on the AVLT (Davis, Millis, & Axelrod, 2012) also showed adequate specificity in the cognitively intact group but not MCI or dementia groups in which false positive rates were 44% and 79%, respectively. One approach to reducing the false positive rate in cases involving suspected dementia is to adjust the cutoffs to achieve adequate specificity (Dean et al., 2009; Loring et al., 2016).

Other research has identified embedded PVTs that appear to have adequate specificity without adjustment (Bortnik, Horner, & Bachman, 2013; Merten et al., 2007), and combinations of PVTs have been found to reduce false positive rates (Loring et al., 2016).

Another area of debate over embedded PVTs and false positive rates centers on the relationship between the number of PVTs administered and PVTs failed. Given the proliferation of validity research and development of embedded PVTs in virtually all cognitive domains (Martin, Schroeder, & Odland, 2015), it is possible for one neuropsychological measure to contribute to multiple embedded PVTs. The Trail Making Test (TMT; Partington & Leiter, 1949; Reitan, 1955), for example, might be used to derive five embedded PVTs: TMT-A and -B raw scores (Iverson, Lange, Green, & Franzen, 2002); TMT-B to -A ratio (Iverson et al., 2002; Merten et al., 2007); and errors on TMT-A and -B (Ruffolo, Guilmette, & Willis, 2000). Although PVT research has continued to prioritize specificity over sensitivity at the level of individual test development (Larrabee & Berry, 2007), some authors have noted concern that the cumulative false positive rate across a test battery might not remain at 10% given the number of PVTs under consideration and likely correlations among measures (Berthelson et al., 2013).

In contrast to research supporting the benefits to classification accuracy arising from use of multiple PVTs (Larrabee, 2003; Loring et al., 2016), Berthelson and colleagues (2013) proposed that excessive false positives would result because the binomial probability theory underlying aggregation across multiple PVTs is premised on independence of PVTs, which is not reflected in actual data (p. 910). They proposed Monte Carlo simulation as an approach to estimation of false positive rates that takes into account the correlations among PVTs. In their metaanalysis of 18 PVT studies, Berthelson and colleagues noted an average PVT intercorrelation of .31 (range .14 - .92) and suggested that PVT false positive rates increase linearly in relation to the number of PVTs administered with an average increase in false positive rate of about 3% and 5% with every additional PVT administered at specificities of .90 and .85, respectively (Tables IV and V in Berthelson et al., 2013).

Issues have been noted with the analytic methods used in the Berthelson et al. (2013) study (Davis & Millis, 2014; Larrabee, 2014; Tarescavage & Ben Porath, 2015). Research based on actual clinical and forensic data has demonstrated improved classification accuracy using multiple PVTs (Davis & Millis, 2014; Larrabee, 2003; Miller et al., 2011; Vickery et al., 2004; Victor, Boone, Serpa, Buehler, & Ziegler, 2009). Proponents of statistical simulation, however, continue to claim a strong linear relationship between PVTs administered and PVTs failed (Odland, Lammy, Martin, Grote, and Mittenberg, 2015).

Considering the continued importance of minimizing false positive rates on PVTs, noted concerns over the use of multiple embedded PVTs, and remaining questions about the utility of embedded PVTs in certain cases (e.g., suspected dementia), further research examining false positive rates on embedded PVTs in cases at risk of or involving dementia is warranted. In addition, predicted false positive rates based on simulated data reported in prior research (Berthelson et al., 2013), warrant empirical cross-validation. Earlier studies undertaking such empirical investigation (Davis & Millis, 2014; Larrabee, 2014) were criticized for small sample size and failure to include results based on null hypothesis testing (Bilder, Sugar, &

Hellemann, 2014). To this end, a large sample comprised of older adults including cognitively normal and impaired groups offers a compelling test for examining the relationship between the number of PVTs administered and PVTs failed. Clinical groups predisposed to high false positive rates like older adults and individuals diagnosed with dementia may show an even greater effect of the number of PVTs administered on PVT failure. Given the claims of earlier research using simulated data (Berthelson et al., 2013; Odland et al., 2015), this study sought to examine the hypothesis that PVT failure rates observed empirically can be accurately predicted by Monte Carlo simulation taking into account the number of PVTs administered and PVT intercorrelations.

Method

Participants

This study involved secondary analysis of data from the National Alzheimer's Coordinating Center (NACC) database. The NACC database is described in detail elsewhere (Beekly, et al., 2007; Weintraub et al., 2009). Briefly, the NACC aggregates and disseminates data gathered by 39 Alzheimer's Disease Centers (ADCs) receiving funding from the National Institute on Aging. The ADCs contribute demographic, diagnostic, neurologic, neuropsychologic, and neuropathologic data from multiple clinic visits. The NACC moved to a Uniform Data Set (UDS) in 2005. This study used data from 35 ADCs. Participants included in this study were seen at NACC sites from 2005 to 2015 and were examined using version 1 and version 2 of the NACC UDS.

The dataset was restricted to the initial visit resulting in 24,127 cases. Cases were excluded for missing data on health history ($n = 893$), neuropsychological variables ($n = 274$), and demographics ($n = 272$) resulting in a final sample of 22,688. The sample was 57.2% female. In terms of race and ethnicity, the sample was 77.6% white, 13.8% black or African American, 6.0% Hispanic/Latino, 2% Asian, 0.4% American Indian or Alaska Native, 0.1% Native Hawaiian or Other Pacific Islander, and 0.1% other. The sample was 72.0 years old on average ($SD = 10.3$). The average educational level was 15.2 years ($SD = 3.1$). Participants were grouped by diagnosis with 45.5% classified as having normal cognition, 25.4% diagnosed with dementia, 23.9% diagnosed with amnesic or non-amnesic mild cognitive impairment (MCI), and 5.3% diagnosed with cognitive impairment but not MCI. Diagnosis was determined clinically at individual ADCs according to published diagnostic criteria based on information captured in the UDS (Beekly, et al., 2007; McKann et al., 1984; Petersen & Morris, 2005; Weintraub et al., 2009). Individuals considered cognitively impaired but not meeting diagnostic criteria for dementia or MCI were classified in the impaired but not MCI group (NACC, 2008). Diagnosis was determined by a single clinician in 22.1% of the sample ($n = 5,025$) and by consensus in 77.9 ($n = 17,663$). UDS version 1 was completed by 46.8% of the sample, and 53.2% completed UDS version 2.

Measures

Participants underwent neuropsychological evaluation with a battery that included common elements according to the UDS covering cognitive domains of attention, language, memory, and executive function (Weintraub et al., 2009). Tests comprising the UDS include a

modified version of the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983), Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), Semantic Word Generation (animal naming and vegetable naming; Gladsjo et al., 1999), Trail Making Test (Partington & Leiter, 1949; Reitan, 1955), Wechsler Adult Intelligence Scale-Revised Digit Symbol Coding (Wechsler, 1987a), and Wechsler Memory Scale-Revised subtests (Digit Span and Logical Memory Story A; Wechsler, 1987b). Given the aim of this study, the UDS neuropsychological test battery was examined for measures identified in previous research with potential for use as embedded PVTs. Measures of interest included Digit Span (DS; Wechsler, 1987a, 1987b), Digit Symbol Coding (COD; Wechsler, 1987a), Logical Memory Story A (LM; Wechsler, 1987b), semantic word generation (SWG; animal naming; Gladsjo et al., 1999), and the Trail Making Test (TMT; Partington & Leiter, 1949; Reitan, 1955).

Embedded PVTs were derived for all participants based on five neuropsychological tests: COD, DS, LM, SWG, and TMT. Cutoffs and references for the embedded PVTs are shown in Table 1. COD raw scores were converted to dichotomous pass or fail score based on an age-adjusted scaled score cutoff (< 6 ; Inman & Berry, 2002). Available DS raw scores included forward trials correct (DSF), longest DS forward, backward trials correct (DSB) and longest DS backward. DSF and DSB scores were combined to produce a total DS raw score that was converted to dichotomous pass or fail score based on an age-adjusted scaled score cutoff (< 6 ; Iverson & Franzen, 1994) using WAIS-R (Wechsler, 1987a) normative data. LM raw scores for Story A were doubled to produce an estimated LM raw score that was examined using a previously identified cutoff (< 14 ; Iverson & Franzen, 1996). SWG raw scores from animal naming were used (< 13 ; Sugarman & Axelrod, 2015). The TMT ratio (TMT-R) was calculated by dividing TMT-B raw score by TMT-A raw score (< 1.5 ; Iverson et al., 2002).

Results

Participant demographics and performance on neuropsychological measures by diagnostic group are shown in Table 2. Given the sample size, all demographic comparisons were significantly different despite small effect sizes in many cases. Performance on neuropsychological measures varied by diagnostic group consistent with what would be expected clinically.

False positive rates on PVTs ranged from a low of 3.3 on TMT to a high of 26.3% on LM in the whole sample. False positive rates by diagnostic group and PVT are shown in Table 3. The average PVT intercorrelation was .40 (range $-.25$ to $.59$; Table 4).

Predicted failure rates based on statistical simulation (Berthelson et al., 2013) were compared to observed failure rates by diagnostic group and assumed specificity. Assuming specificity of .90 and failure on 2 or more of 5 PVTs, Berthelson and colleagues predicted a false positive rate of 11.5%. In the normal group, observed (1.9%) and predicted rates were significantly different, $z = 30.66$, $p < .0001$. In the impaired group, observed (6.6%) and predicted rates were significantly different, $z = 5.29$, $p < .0001$. In the MCI group, observed (13.2%) and predicted rates were significantly different, $z = 3.97$, $p = .0001$. In the dementia

group, observed (52.8%) and predicted rates were significantly different, $z = 98.19$, $p < .0001$.

Assuming specificity of .85 and failure on 2 or more of 5 PVTs, Berthelson and colleagues (2013) predicted a false positive rate of 19.4%. Observed and predicted proportions were again examined by diagnostic group. In the normal group, observed (1.9%) and predicted rates were significantly different, $z = 45.03$, $p < .0001$. In the impaired group, observed (6.6%) and predicted rates were significantly different, $z = 11.17$, $p < .0001$. In the MCI group, observed (13.2%) and predicted rates were significantly different, $z = 11.50$, $p < .0001$. In the dementia group, observed (52.8%) and predicted rates were significantly different, $z = 64.05$, $p < .0001$.

Observed and predicted false positive rates by diagnostic group and specificity are shown in Figure 1.

During peer review, additional analyses were suggested to compare observed and predicted false positive rates with failure on 3 or more of 5 PVTs. Berthelson et al. (2013) predicted false positive rates of 3.6% and 7.3% with specificity of .90 and .85, respectively, and failure on 3 or more of 5 PVTs. The observed false positive rates in the normal (0.17%), impaired (1.5%), and MCI (1.4%) groups were significantly lower than predicted false positive rates at both specificity levels ($z = 3.9$ to 18.7 , $p < .0001$). In the dementia group, the observed false positive rate (13.3%) was significantly higher than predicted rates at both specificity levels ($z = 17.5$ to 39.5 , $p < .0001$).

Discussion

This study examined false positive rates on embedded PVTs administered to an older adult sample categorized into cognitively normal, impaired, MCI, and dementia groups. Despite examination of 5 embedded PVTs, the average false positive rate in the cognitively normal group remained within a commonly accepted range (Larrabee & Berry, 2007) with all measures showing false positive rate less than 10%. In the impaired and MCI groups, false positive rates were acceptable on three of five measures (COD, DS, and TMT-R). In the dementia group, the false positive rate remained acceptable on TMT-R. These findings are consistent with prior research in clinical samples with diagnosed memory disorders (Dean et al., 2009; Loring et al., 2016).

These observations indicate that a number of embedded PVTs maintain adequate specificity in older adults without diagnosed cognitive impairment. Therefore, in cases involving no or minimal expectation of neurologic cognitive dysfunction (e.g., three or more months after mild traumatic brain injury), older age alone is not a sufficient reason to consider PVT failure a false positive. Another observation from these findings is that, despite the clear increase in false positive rates seen with increasing cognitive dysfunction (i.e., a tenfold increase in the proportion failing two or more PVTs from cognitively normal to dementia groups), several embedded PVTs maintained adequate specificity in all but the dementia group. TMT-R demonstrated adequate specificity even in the dementia group. COD cutoffs have previously been examined only in simulation and mild traumatic brain injury samples

(Inman & Berry, 2002). DS variables appeared promising in some research involving dementia (Dean et al., 2009) but required adjustment to maintain adequate specificity in another sample (Loring et al., 2016). TMT variables appeared useful in prior research in samples involving memory deficits and dementia (Bortnik et al., 2013; Merten et al., 2007). Current findings thus extend prior research on use of embedded PVTs in cases involving suspected or diagnosed memory deficits including dementia.

A related implication of these observations is that in cases involving mild traumatic brain injury or another condition with limited expectation of significant cognitive compromise, individuals performing in the range of or lower than clinical groups with diagnosed cognitive deficits are likely performing noncredibly. Current findings might prove useful in contextualizing such performances. For example, a litigant claiming cognitive deficits following mild traumatic brain injury who completes TMT-A in 62 seconds or more, or completes TMT-B in 212 seconds or more can be described as performing in the range of older adults diagnosed with dementia. If such an individual drives him- or herself to the appointment, arrives on time, and navigates the sometimes challenging visuospatial demands of modern medical facilities, then it appears highly implausible that the TMT performance is an accurate reflection of cognitive dysfunction with a neurologic basis.

Regarding the proposed relationship between PVT administration and PVT failure (Berthelson, et al., 2013; Odland et al., 2015), the current findings provide striking evidence contrary to what has been predicted based on simulated data. Statistical comparison of observed and predicted false positive rates at different specificity levels (i.e. .90 or .85) revealed significant differences in all comparisons. Predicted false positive rates based on simulation demonstrated minimal relationship to observed false positive rates with overestimation in cognitively normal older adults as well as among older adults with a diagnosis of impairment, underestimation in individuals diagnosed with dementia, and variable findings in the MCI group. The instances in which observed rates were higher than predicted rates occurred in participants with a diagnosis of MCI or dementia, which have well-known susceptibility to false positive errors on PVTs. These findings do not support the utility of predicted false positive rates based on simulated data in a sample of older adults grouped by diagnosis. As noted in practice guidelines (Heilbrunner et al., 2009), interpretation of PVT findings should occur in consideration of clinical characteristics observed in a specific case not in response to overreaching recommendations based on simulated data.

Limitations of this study include restriction in the range of neuropsychological data to variables included in the UDS, which incorporates older versions of tests to maintain consistency over time (Weintraub et al., 2009). Therefore, observed false positive rates on individual tests may be less relevant for those tests that have undergone revision (i.e., addition of new items to DS and COD subtests). Also, data were only available for LM Story A necessitating the use of an estimated LM raw score, which may underestimate actual LM raw score potentially inflating the false positive rate on the LM cutoff. Another limitation arises from the lack of a gold standard criterion for categorizing participants in terms of validity. Given the goal of the current study, it was thought to be more conservative to consider all PVT failures as false positives. While this approach may over-estimate the

false positive rate, it appears a stringent test of the proposed association between the number PVTs administered and PVT failure. Finally, the sample was well-educated, which may limit extension of these findings to populations with lower educational attainment.

Future research might extend these findings by examining cutoffs on updated versions of instruments in the UDS. Given the promising findings of preserved specificity with some embedded PVTs, further research might also examine optimal cutoffs for different diagnostic groups (i.e., healthy older adults versus individuals with MCI or dementia diagnoses) on these measures. Furthermore, aggregation of additional information on relationships (or lack thereof) between the number of PVTs administered and PVTs failed would be facilitated if future validity studies were to report the number of PVTs administered, the number of PVTs failed, and, if multiple PVTs were administered, the non-parametric correlation coefficient between PVTs administered and failed. Given the ongoing productivity in PVT research, the addition of a few simple analyses would support future meta-analytic efforts that ultimately might provide stronger evidence of factors influencing PVT failure than what can be offered by any single study regardless of sample size.

Acknowledgments

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD), P50 AG016574 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI M. Marsel Mesulam, MD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P50 AG005131 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG033514 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

References

- Babikian T, Boone KB, Lu P, Arnold G. Sensitivity and specificity of various Digit Span scores in the detection of suspect effort. *The Clinical Neuropsychologist*. 2006; 20:145–159. [PubMed: 16393925]
- Beekly DL, Ramos EM, Lee WW, Deitrich WD, Jacka ME, Wu J, Kukull WA. The National Alzheimer's Coordinating Center (NACC) database: The uniform data set. *Alzheimer Disease & Associated Disorders*. 2007; 21:249–258. [PubMed: 17804958]
- Berthelson L, Mulchan SS, Odland AP, Miller LJ, Mittenberg W. False positive diagnosis of malingering due to the use of multiple effort tests. *Brain Injury*. 2013; 27:909–916. [PubMed: 23782260]
- Bilder RM, Sugar CA, Helleman GS. Cumulative false positive rates given multiple performance validity tests: Commentary on Davis and Millis (2014) and Larrabee (2014). *The Clinical Neuropsychologist*. 2014; 28:1212–1223. [PubMed: 25490983]
- Boone KB. The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist*. 2009; 23:729–741. [PubMed: 18949583]
- Bortnik KE, Horner MD, Bachman DL. Performance on standard indexes of effort among patients with dementia. *Applied Neuropsychology: Adult*. 2013; 20:233–242.

- Bush SS, Ruff RM, Troster AI, Barth JT, Koffler SP, Pliskin NH, Silver CH. Symptom validity assessment: Practice issues and medical necessity – NAN policy & planning committee. *Archives of Clinical Neuropsychology*. 2005; 20:419–426. [PubMed: 15896556]
- Davis JJ, Millis SR. Examination of performance validity test failure in relation to number of tests administered. *The Clinical Neuropsychologist*. 2014; 28:199–214. [PubMed: 24528190]
- Davis JJ, Millis SR, Axelrod BN. Derivation of an embedded Rey Auditory Verbal Learning Test performance validity indicator. *The Clinical Neuropsychologist*. 2012; 26:1397–1408. [PubMed: 23075400]
- Dean AC, Victor TL, Boone KB, Arnold G. The relationship of IQ to effort test performance. *The Clinical Neuropsychologist*. 2008; 22:705–722. [PubMed: 17853124]
- Dean AC, Victor TL, Boone KB, Philpott LM, Hess RA. Dementia and effort test performance. *The Clinical Neuropsychologist*. 2009; 23:133–152. [PubMed: 18609332]
- Duff K, Spering CC, O'Bryant SE, Beglinger LJ, Moser DJ, Bayless JD, Scott JG. The RBANS Effort Index: Base rates in geriatric samples. *Applied Neuropsychology*. 2011; 18(1):11–17. [PubMed: 21390895]
- Etherton JL, Bianchini KJ, Heinly MT, Greve KW. Pain, malingering, and performance on the WAIS-III Processing Speed Index. *Journal of Clinical and Experimental Neuropsychology*. 2006; 23:1218–1237.
- Faust D, Guilmette TJ. To say it's not so doesn't prove that it isn't: Research on the detection of malingering. Reply to Bigler. *Journal of Consulting and Clinical Psychology*. 1990; 58:248–250. [PubMed: 2335643]
- Folstein MF, Folstein SE, McHugh PR. Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*. 1975; 12:189–198. [PubMed: 1202204]
- Gladsjo JA, Schuman CC, Evans JD, Peavy GM, Miller SW, Heaton RK. Norms for letter and category fluency: Demographic corrections for age, education, and ethnicity. *Assessment*. 1999; 6:147–160. [PubMed: 10335019]
- Greiffenstein MF, Baker WJ, Gola T. Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*. 1994; 6:218–224.
- Heaton RK, Miller SW, Taylor MJ, Grant I. Revised comprehensive norms for an expanded Halstead-Reitan battery: Demographically adjusted neuropsychological norms for African Americans and Caucasian adults (professional manual). Lutz, FL: Psychological Assessment Resources; 2004.
- Heilbrunner RL, Sweet JJ, Morgan JE, Larrabee GJ, Millis SR, Conference Participants. American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*. 2009; 23:1093–1129. [PubMed: 19735055]
- Heinly MT, Greve KW, Bianchini KJ, Love JM, Brennan A. WAIS Digit Span-based indicators of malingered neurocognitive dysfunction: Classification accuracy in traumatic brain injury. *Assessment*. 2005; 12:429–444. [PubMed: 16244123]
- Howe LS, Loring DW. Classification accuracy and predictive ability of the Medical Symptom Validity Test's Dementia Profile and General Memory Impairment Profile. *The Clinical Neuropsychologist*. 2009; 23(2):329–342. [PubMed: 18609326]
- Inman TH, Berry DT. Cross-validation of indicators of malingering: A comparison of nine neuropsychological tests, four tests of malingering, and behavioral observations. *Archives of Clinical Neuropsychology*. 2002; 17:1–23. [PubMed: 14589749]
- Iverson GL, Franzen MD. The Recognition Memory Test, Digit Span, and Knox Cube Test as markers of malingered memory impairment. *Assessment*. 1994; 1:323–334.
- Iverson GL, Franzen MD. Using multiple objective memory procedures to detect simulated malingering. *Journal of Clinical and Experimental Neuropsychology*. 1996; 18:38–51. [PubMed: 8926295]
- Iverson GL, Lange RT, Green P, Franzen MD. Detecting exaggeration and malingering with the Trail Making Test. *The Clinical Neuropsychologist*. 2002; 16:398–406. [PubMed: 12607151]

- Ivnik RJ, Malec JF, Smith GE, Tangalos EG, Petersen RC, Kokmen E, Kurland LT. Mayo's older Americans normative studies: Updated AVLT norms for ages 56–97. *The Clinical Neuropsychologist*. 1992; 6:83–104.
- Kaplan E, Goodglass H, Weintraub S. *The Boston Naming Test*. Philadelphia: Lea and Febiger; 1983.
- Kiewel NA, Wisdom NM, Bradshaw MR, Pastorek MJ, Strutt AM. A retrospective review of Digit Span-related effort indicators in probable Alzheimer's disease patients. *The Clinical Neuropsychologist*. 2012; 26:965–974. [PubMed: 22703555]
- Larrabee GJ. Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist*. 2003; 17:410–425. [PubMed: 14704892]
- Larrabee GJ. Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*. 2012; 18:625–631. [PubMed: 23057079]
- Larrabee GJ. False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology*. 2014; 29:364–373. [PubMed: 24769887]
- Larrabee GJ, Berry DTR. Diagnostic Classification Statistics and Diagnostic Validity of Malingering Assessment. In: Larrabee GJ, editor *Assessment of malingered neuropsychological deficits*. New York: Oxford; 2007. 14–26.
- Loring DW, Goldstein FC, Chen C, Drane DL, Lah JJ, Zhao L, Larrabee GJ. 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*. 2016; 31:313–331. [PubMed: 27084732]
- Martin PK, Schroeder RW, Odland AP. Neuropsychologists' validity testing beliefs and practices: A survey of North American professionals. *The Clinical Neuropsychologist*. 2015; 29:741–776. [PubMed: 26390099]
- McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology*. 1984; 34:939–944. [PubMed: 6610841]
- Merkelbach H, Jelicic M, Pieters M. The residual effect of feigning: How intentional faking may evolve into a less conscious form of symptom reporting. *Journal of Clinical and Experimental Neuropsychology*. 2011; 33:131–139. [PubMed: 20623399]
- Merten T, Bossink L, Schmand B. On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology*. 2007; 29:308–318. [PubMed: 17454351]
- Miller JB, Millis SR, Rapport LJ, Bashem JR, Hanks RA, Axelrod BN. Detection of insufficient effort using the advanced clinical solutions for the Wechsler Memory Scale, fourth edition. *The Clinical Neuropsychologist*. 2011; 25:160–172. [PubMed: 21253964]
- National Alzheimer's Coordinating Center. NACC Uniform Data Set (UDS) coding guidebook for initial visit packet. 2008. Retrieved from https://www.alz.washington.edu/WEB/forms_uds.html
- Odland A, Lammy A, Martin P, Grote C, Mittenberg W. Advanced administration and interpretation of multiple validity tests. *Psychological Injury and Law*. 2015; 8:46–63.
- Partington JE, Leiter RG. Partington's pathways test. *The Psychological Services Center Bulletin*. 1949; 1:9–20.
- Petersen RC, Morris JC. Mild cognitive impairment as a clinical entity and treatment target. *Archives of Neurology*. 2005; 62:1160–1163. [PubMed: 16009779]
- Reitan RM. The relation of the Trail Making Test to organic brain damage. *Journal of Consulting Psychology*. 1955; 19:393–394. [PubMed: 13263471]
- Rey A. L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de Psychologie*. 1941; 28:286–340.
- Ruffolo LF, Guilmette TJ, Willis GW. FORUM: Comparison of time and error rates on the Trail Making Test among patients with head injuries, experimental malingerers, patients with suspect effort on testing, and normal controls. *The Clinical Neuropsychologist*. 2000; 14:223–230. [PubMed: 10916197]

- Salazar XF, Lu PH, Wen J, Boone KB. The use of effort tests in ethnic minorities and in non-English-speaking and English as a second language populations. In: Boone KB, editor *Assessment of feigned cognitive impairment*. New York: Guilford; 2007. 405–427.
- Sharland M, Gfeller J. A survey of neuropsychologists' beliefs and practices with respect to the assessment of effort. *Archives of Clinical Neuropsychology*. 2007; 22:213–224. [PubMed: 17284353]
- Sugarman MA, Axelrod BN. Embedded measures of performance validity using verbal fluency tests in a clinical sample. *Applied Neuropsychology: Adult*. 2015; 22:141–146. [PubMed: 25153155]
- Sweet JJ, Guidotti Breting LM. Symptom validity test research: Status and clinical implications. *Journal of Experimental Psychopathology*. 2013; 4:6–19.
- Tarescavage AM, Ben-Porath YS. A response to Odland et al.'s misleading, alarmist estimates of risk for overpathologizing when interpreting the MMPI-2-RF. *The Clinical Neuropsychologist*. 2015; 29:183–196. [PubMed: 25947107]
- Teichner G, Wagner MT. The Test of Memory Malinger (TOMM): Normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Archives of Clinical Neuropsychology*. 2004; 19:455–464. [PubMed: 15033228]
- Tombaugh T. *Test of memory malinger*. North Tonawanda, NY: Multi-Health Systems; 1996.
- Van Dyke SA, Millis SR, Axelrod BN, Hanks RA. Assessing effort: Differentiating performance and symptom validity. *The Clinical Neuropsychologist*. 2013; 27:1234–1246. [PubMed: 24028487]
- Vickery CE, Berry DTR, Dearth CS, Vagnini VL, Baser RE, Cragar DE, Orey SA. Head injury and the ability to feign neuropsychological deficits. *Archives of Clinical Neuropsychology*. 2004; 19:37–48. [PubMed: 14670378]
- Victor TL, Boone KB, Serpa JG, Buehler J, Ziegler EA. Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist*. 2009; 23:297–313. [PubMed: 18821138]
- Webb JW, Batchelor J, Meares S, Taylor A, Marsh NV. Effort test failure: Toward a predictive model. *The Clinical Neuropsychologist*. 2012; 26(8):1377–1396. [PubMed: 23061431]
- Wechsler D. *Wechsler Adult Intelligence Scale-Revised*. San Antonio, Texas: Psychological Corporation; 1987a.
- Wechsler D. *Wechsler Memory Scale-Revised Manual*. San Antonio, TX: Psychological Corporation; 1987b.
- Whiteside DM, Kogan J, Wardin L, Phillips D, Franzwa MG, Rice L, Basso M, Roper B. Language-based embedded performance validity measures in traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*. 2015; 37:220–227. [PubMed: 25655924]

Figure 1.

Observed (bars) false positive rate by diagnostic group and predicted (lines) false positive rates by specificity with failure on 2 or more of 5 PVTs. All comparisons between observed and predicted false positive rates were significantly different ($p < .0001$). Predicted rates obtained from Berthelson et al. (2013). PVT = performance validity test; Pre-85 = Predicted false positive rate assuming 85% specificity; Pre-90 = Predicted false positive rate assuming 90% specificity; NOR = cognitively normal; IMP = impaired; MCI = amnesic or non-amnesic mild cognitive impairment; DEM = dementia.

Table 1

Embedded performance validity test cutoffs, classification accuracy, and references

Measure	Cutoff	SEN	SPE	References
COD AASS	< 6	.02 - .81	.62 - .87	Inman & Berry, 2002; Etherton et al., 2006
DS AASS	< 6	.25 - .73	.22 - 1.0	Babikian et al., 2006; Dean et al., 2009; Iverson & Franzen, 1994; Heinly et al., 2005; Kiewel et al., 2012
LM	< 14	.60	.95	Iverson & Franzen, 1996
SWG	< 13	.25 - .49	.89 - .90	Sugarman & Axelrod, 2015; Whiteside et al., 2015
TMT Ratio	< 1.5	.02 - .07	.87 - .96	Iverson et al., 2002; Merten et al., 2007

Note. SEN = range of published sensitivities at cutoff; SPE = range of published specificities at cutoff; COD = Digit-Symbol Coding; AASS = age-adjusted scaled score; DS = Digit Span; LM = Logical Memory I estimated raw score (based on doubling Story A raw score); SWG = semantic word generation raw score; TMT-Ratio = Trail Making Test-B to -A ratio.

Table 2

Participant demographics and neuropsychological performance by diagnostic group

Variable	Normal (<i>n</i> = 10,319)	Impaired (<i>n</i> = 1,194)	MCI (<i>n</i> = 5,414)	Dementia (<i>n</i> = 5,761)
Age	70.6 (10.7) _a	70.4 (10.1) _a	73.2 (9.4) _b	73.5 (9.8) _b
% Female	65.8%	56.0%	49.5%	49.4%
% Caucasian	77.2%	69.4%	75.2%	82.4%
Education	15.6 (2.9) _a	15.0 (3.3) _b	15.1 (3.3) _b	14.7 (3.3) _c
COD	47.8 (12.6) _a	41.7 (12.5) _b	37.3 (12.0) _c	27.8 (13.2) _d
DS	15.4 (3.8) _a	14.1 (3.8) _b	13.6 (3.6) _c	11.9 (3.6) _d
LM	26.9 (7.7) _a	23.1 (8.5) _b	18.4 (8.3) _c	10.3 (7.7) _d
SWG	20.2 (5.6) _a	18.2 (5.3) _b	16.0 (5.0) _c	11.6 (5.0) _d
TMT-A	34.3 (15.9) _a	40.2 (21.1) _b	44.3 (22.3) _c	62.2 (35.2) _d
TMT-B	90.8 (51.5) _a	119.7 (69.9) _b	142.1 (78.2) _c	212.3 (88.1) _d
TMT-R	2.7 (1.1) _a	3.1 (1.4) _b	3.3 (1.5) _c	3.9 (1.9) _d

Note. Values are *M*(*SD*) except as noted. All group comparisons were significant ($p < .0001$). Cells with different subscripts were significantly different on post hoc analyses (Sidak) conducted on continuous variables. MCI = mild cognitive impairment; COD = Digit Symbol Coding raw score; DS = Digit Span raw score; LM = Logical Memory I estimated raw score (based on doubling Story A raw score); SWG = semantic word generation raw score; TMT = Trail Making Test; A = TMT-A raw score; B = TMT-B raw score; R = TMT-B to -A ratio.

Table 3

False positive rates by performance validity test and diagnosis

PVT	Normal (<i>n</i> = 10,319)	Impaired (<i>n</i> = 1,194)	MCI (<i>n</i> = 5,414)	Dementia (<i>n</i> = 5,761)	Total (<i>N</i> = 22,688)
COD	0.6	2.4	2.6	16.1	5.1
DS	1.2	3.9	4.3	11.0	4.6
LM	3.8	12.1	27.3	68.4	26.3
SWG	6.9	13.1	25.0	59.8	24.9
TMT-R	4.4	3.0	2.8	1.7	3.3
Average	3.4	6.9	12.4	31.4	12.8

Note. Values are percentages. PVT = performance validity test; MCI = mild cognitive impairment; COD = Digit Symbol Coding raw score; DS = Digit Span raw score; LM = Logical Memory I estimated raw score (based on doubling Story A raw score); SWG = semantic word generation raw score; TMT-R = Trail Making Test-B to -A ratio.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Correlations among performance validity tests

	COD	DS	LM	SWG	TMT-R
COD	–				
DS	.45	–			
LM	.53	.40	–		
SWG	.59	.42	.56	–	
TMT-R	–.26	–.28	–.31	–.25	–

Note. All correlations significant ($p < .0001$). COD = Digit Symbol Coding raw score; DS = Digit Span raw score; LM = Logical Memory I estimated raw score (based on doubling Story A raw score); SWG = semantic word generation raw score; TMT-R = Trail Making Test-B to -A ratio.