Title

Cumulative False Positive Rates Given Multiple Performance Validity Tests: Commentary on Davis and Millis (2014) and Larrabee (2014)

Journal
The Clinical Neuropsychologist, 28(8)

ISSN
1385-4046

Authors
Bilder, Robert M
Sugar, Catherine A
Hellemann, Gerhard S

Publication Date
2014-11-17

DOI
10.1080/13854046.2014.969774

Peer reviewed

Published in final edited form as:
Clin Neuropsychol. 2014 November ; 28(8): 1212–1223. doi:10.1080/13854046.2014.969774.

Cumulative False Positive Rates Given Multiple Performance Validity Tests: Commentary on Davis and Millis (2014) and Larrabee (2014)

Robert M. Bilder,
Department of Psychiatry & Biobehavioral Sciences, David Geffen School of Medicine at UCLA, and Department of Psychology, UCLA College of Letters & Science

Catherine A. Sugar, and
Department of Biostatistics, UCLA School of Public Health, and Department of Psychiatry & Biobehavioral Sciences, David Geffen School of Medicine at UCLA

Gerhard S. Hellemann
Semel Institute for Neuroscience & Human Behavior at UCLA

Abstract

Controversy has arisen over interpretation of performance validity tests (PVTs) when multiple PVTs are given. Some papers state that more stringent criteria are needed to judge overall performance as invalid, while others argue that concerns about the number of PVTs are overstated and that widely used criteria are appropriate. We examine theoretical models and assumptions, and analyze published data to determine the magnitude of effects implied by theory and observed in practice. Assertions advanced in the primary papers are examined for consistency with the empirical data. Existing theoretical models do not account well for the diverse empirical data, substantial empirical effects remain poorly understood, and the primary papers include assertions that are not empirically supported. The results indicate that: (a) neuropsychology lacks solid theoretical bases for estimating PVT failure rates given various combinations of PVTs, and thus needs to rely on empirical data; (b) existing empirical data fail to support the application of any uniform criteria across the broad range of scenarios involving multiple PVTs; and (c) practice should rely on empirical studies involving combinations of PVTs that have been studied together, in samples clearly appropriate to the individual case, using experimental designs germane to the questions under consideration.

Keywords

performance validity; symptom validity; forensic; medico-legal; evidence-based

Clinical neuropsychology has focused much attention recently on performance validity tests (PVTs), which are often used to judge the credibility of results from entire test batteries, and have a major effect on high-stakes outcomes, especially in forensic contexts. Several recent

Correspondence concerning this article should be addressed to Robert M. Bilder, Semel Institute for Neuroscience & Human Behavior, 740 Westwood Plaza, Los Angeles, CA 90095. rbilder@mednet.ucla.edu.

publications have generated controversy about the criteria used to determine that overall performance on a battery of neuropsychological tests is invalid, based on results from multiple PVTs. One recent paper based on simulations suggested that widely used criteria for invalidity (e.g., two failed PVTs) may yield unacceptably high false positive rates (FPRs) if the criteria neglect to consider the number of PVTs administered and their inter-correlations (Berthelson, Mulchan, Odland, Miller, & Mittenberg, 2013). A subsequent report in The Clinical Neuropsychologist considered empirical data, which were claimed to "mitigate concerns" about multiple PVTs inflating FPRs (Davis & Millis, 2014). Following this, another report presented empirical data and interpreted these as supporting continued use of the "two failed PVTs" criterion; it was asserted further that adding PVTs does not have a significant impact on the FPR up through 7 PVTs administered (Larrabee, 2014). Given the centrality of these issues to forensic neuropsychology practice, together with the importance of the outcomes for the parties involved in litigation, we aimed to examine the theoretical assumptions and empirical data at the heart of this controversy.

Assumptions: Can We Predict PVT Failure from Theoretical Models?

Brief background helps put the controversy in perspective. Initial attempts to determine appropriate cutoffs for invalidity on multiple PVTs were based on theoretical probabilistic models of false positive rates; this work has recently been expanded via simulations and models fit using empirical data. We briefly review the models that have been proposed.

Independent Binomial Model

Early work suggested that an "…overall false positive rate could be obtained by multiplying the individual false positive rates together" (Boone and Lu, 2003; p. 252). For example, the likelihood of failing 6 PVTs, each with a false positive rate of 10%, was said to be "one in a million" ($.10^6$)(Boone and Lu; page 252). This approach was generalized to consider the overall probability of failing a *specific number of tests given the number of PVTs administered*. Assuming standard binomial probability distributions and tests with individual false positive rates of 10%[1], the theoretical likelihood of failing 2 of 5 PVTs by chance is 7.3% and that of failing 3 of 5 PVTs is 0.8% (Larrabee, 2008). These calculations are all based on the assumption that the individual PVT results are independent and that the tests have the same false positive rate.

Correlated Binomial Model

Berthelson and colleagues (2013) reviewed 22 studies, which collectively showed that PVTs are generally not independent. They therefore conducted Monte Carlo simulations to calculate the expected FPRs under a correlated binomial failure model. Specifically, the simulations assumed normally distributed PVT scores with uniform correlations between all pairs of tests, using the mean inter-correlation found among PVTs in their review. The simulations showed unacceptably high FPRs using standard criteria. For example, the simulations indicated that using the 2-failed-PVT criterion, the overall FPR would exceed

[1] Unless otherwise specified, all scenarios presented here assume that the false positive rate for each individual PVT is 10% (or that "specificity" is 90%). Berthelson and colleagues noted that actual FPRs reported for individual tests in practice sometimes exceed 15%, which further increases estimates of the overall FPRs (see Berthelson et al., 2013, Table V).

10% in a battery including 5 or more PVTs. To maintain an overall FPR of <10%, they found that the criteria for invalidity should be failure on at least 3 of 5 PVTs, with more stringent criteria if even more PVTs are given (Berthelson et al., 2013; Table IV, p. 913).

Negative Binomial Regression Model

Davis and Millis (2014), due to concerns about the assumptions made by Berthelson and colleagues in their correlated binomial model, proposed using negative binomial regression to determine whether the number of PVTs administered was associated with the number of failed PVTs based on empirical data from a mixed sample of 158 individuals. This approach is thus a hybrid of theoretical modeling and empirical estimation. In addition to the number of PVTs administered, the regression model included age, education, medico-legal context, and activities of daily living (ADL). The only significant predictors were ADL and education (poorer ADL and lower education were both associated with more PVTs failed). The results were interpreted as evidence that the number of PVTs does not have much effect on the overall FPR, and that the number of PVTs given is less important than medicolegal context in determining the likelihood of failing multiple PVTs.

Empirical Data on PVT Failure Rates and Comparisons with the Model-Based Approaches

We consider below the assumptions made by each of the above theoretical models, compare the estimated FPRs from each of those theoretical models with available empirical data, and review the assertions, interpretations, and conclusions drawn about them in the relevant articles. Finally, we discuss the implications of the findings for practice and future research.

Binomial Models

Neither the original independent binomial model, nor the modified binomial model that considers correlations among PVTs, offers a good fit to the diversity of empirical data on PVTs. Larrabee (2014) provided evidence that data from earlier publications (Larrabee 2003, 2009) show lower FPRs than predicted by either independent or correlated binomial models, and that other data (Pearson, 2009; Schroeder & Marshall, 2011; Victor, Boone, Serpa, Buehler, & Ziegler, 2009) have significantly lower FPRs than the Berthelson simulations. On the other hand, Berthelson and colleagues' simulation data are very similar to both the empirical data of Pella and colleagues (2012), and to the empirical data of Davis and Millis (2014), as shown below. We examine details of these comparisons because they raise multiple substantive concerns about both the underlying assumptions and some of the interpretations of the data.

To assess the validity of the correlated binomial model, Davis and Millis (2014) reported empirical data on 87 people in a "neurological no-incentive" (NNI) group (in which PVT failures were assumed to reflect false positives), and compared these data to Berthelson and colleagues' simulation data. The Davis and Millis (2014) paper states:

> "…observed false positive rates were compared to predictions offered by Berthelson et al. (2013)… In almost all cases the observed failure rates were lower than prediction" (page 209); and

> “Observed rates of failing zero PVTs were greater than prediction. Observed proportions of participants failing one or more and two or more PVTs were lower than prediction” (page 206).

But the Davis and Millis paper offers no statistical support for these assertions. We conducted statistical tests of the differences between the Berthelson et al. (2013) data and the empirical data of Davis and Millis (2014) over the range of 6, 7, or 8 PVTs administered (following Davis & Millis, Table 5), and find that the observed and predicted data do not differ significantly for any of these comparisons. Indeed, the number of cases failing two or more PVTs in the Davis and Millis study is almost exactly what Berthelson and colleagues' simulation predicts, within rounding errors (Table 1)[2].

To further determine the consistency of results, we tabulated the predicted failure rates under each of the primary models:

1. The standard binomial model (assuming independence, following Larrabee 2003);
2. The adjusted binomial model (assuming a uniform inter-test correlation of .31, following Berthelson et al., 2013); and
3. The empirical observations of Davis & Millis (2014), Larrabee (2003, 2009), and Pearson (2009) (the latter referred to as “ACS” because these are from Pearson Advanced Clinical Solutions).

Table 2 shows the FPRs from both the theoretical estimates and the empirical data from these studies, at different cutoff criteria for overall invalidity (from 2-or-more through 6-or-more failed PVTs), given the number of PVTs administered (from 5 PVTs to 8 PVTs).

Table 2 highlights FPRs > 10% (darker shading), given widespread agreement that these FPRs are unacceptable. Table 2 also indicates FPRs greater than 5% but less than 10% (lighter shading), given opinions that overall FPRs should be more stringent considering the potential adverse consequences of false positive errors.[3]

Table 2 shows that for the scenarios in which 5 PVTs are given, the criterion of 2 or more failed PVTs yields FPRs under 10% except for the correlated binomial model, the negative binomial model (with 12 years of education), and the Davis-Millis data (but these are based on only 5 cases). In the scenarios involving 6 PVTs, only the negative binomial (with 14 years of education) and the Davis-Millis data yield FPR < 10%. In the scenarios involving 7 or 8 PVTs, the criterion of 2 or more failed PVTs is unacceptable (FPRs > 10%) for all scenarios. For the scenarios involving 8 PVTs, even a criterion of 3 or more failed PVTs is not uniformly supported. This appears to contradict the conclusions of Larrabee (2014) that the “two failed PVT” criterion is adequate to maintain FPR at acceptable levels. Indeed, the Larrabee data suggest that 3 or more PVTs must be failed to maintain overall FPR < 10% when 7 PVTs are administered.

[2] Davis and Millis make 3 different assertions about their findings, but it should be noted that testing the rate of “failing zero PVTs” is the formal complement to the test of rates of “failing one or more PVTs,” so Davis and Millis actually had only two testable hypotheses. Table 1 shows that none of their assertions are supported by their data.

[3] It is beyond the scope of this paper to address multiple additional assumptions. For example, all the arguments here take for granted that the false positive rates are valid for each individual PVT.

Larrabee (2014) advances several other arguments to suggest that the Berthelson et al (2013) simulations of FPRs are biased and overstate the FPRs. First, he suggests that the empirical data (N = 478, college students with no incentive) from Pella and colleagues (Pella, Hill, Shelton, Elliott, & Gouvier, 2012), which are highly consistent with the Berthelson et al simulations, are overestimates. The first argument is that the PVTs used were not really independent, and indeed many were derived from the same underlying tests (e.g., Digit Span). While observed correlations often reflect a combination of influences, including covariance due to shared methods and covariance of true scores on underlying traits, the observed correlations in the Pella et al. paper (r = .32 for non-clinical and r = .38 for clinical samples) are well within the range of values for other studies (see Berthelson et al., Table II, where the mean within-study inter-correlations for PVTs ranged from r = .14 to r = .92, and the mean inter-correlation observed across studies was r = .31). In contrast, the Larrabee (2003; 2009) combined PVT data had an average inter-correlation of r = 0.063, which is *lower* than any of the 22 studies tabulated by Berthelson and colleagues (2013; Table II). To the extent that these 22 studies are representative, the Larrabee results would therefore be expected to yield atypically low estimates of FPR due to the atypically low inter-correlation among PVTs. In contrast, the data of Pella and colleagues are closer to the average of empirical inter-correlations. More important, however, is the fact that the broad range of inter-correlations observed in empirical studies makes it inappropriate to generalize about the correlational structure of PVTs. Observed FPRs are really relevant only for the specific combinations of PVTs studied in a specific empirical study.

The second argument advanced in the Larrabee (2014) paper is that the FPR reported by Pella and colleagues is elevated because it is based on discrepancy scores in higher functioning samples where the “basal” score (in this example, it is Vocabulary, to which other scores are compared) is higher than would be expected in the general population. This is a valuable point, and highlights the fact that we need to consider sample characteristics before interpreting PVTs. Acknowledging this problem may indicate that discrepancy-based PVTs (for example, Vocabulary minus Digit Span, or the “Mittenberg Index”) may only be valid in individuals within a specific range of ability, and that different criteria may be needed at different levels of ability. Even for PVTs that do not involve difference scores, it is important to determine whether a given PVT has been validated, with attention to its sensitivity and specificity, at all relevant levels of ability before interpreting results in a specific case[4].

The Larrabee paper next highlights the findings of Victor et al. (2009) as showing an FPR lower than predicted by Berthelson et al. (2013). Victor et al. (2009), however, used inclusion/exclusion criteria that most likely biased results. First, that paper defined a credible group based not only on lack of incentive, but also based on the criterion that the patients had not failed two or more “freestanding” PVTs (i.e., “stand-alone” PVTs given during the examination), before considering the failure rates on “embedded” PVTs (PVT

[4] The impact of examinee characteristics is seen clearly in the Davis and Millis (2014) negative binomial regression, which highlights how important education level is for estimated overall FPR. Table 2 shows that using 12 rather than 14 years of education shifts the FPR (using the 2 or more failed PVT criterion) from 7.1% up to 11.4% (if 5 PVTs are given), and from 15% to 22% (if 8 PVTs are given). Using the same parameters in Table 2, but dropping education to only 8 years, an examinee would need to fail 4 of 5 PVTs to keep overall FPR < 10%.

indices derived from scores on other ability tests in the battery). If any cases from the original no-incentive sample were excluded based on failing PVTs (the paper does not report the numbers excluded for each criterion), this would have biased the sample to exclude patients who were truly credible, but failed PVTs for reasons unrelated to incentive. This bias eliminates those from the credible group who would be most likely to fail PVTs and may artificially decrease observed FPRs. Second, given that cases assigned to the non-credible group included some who were feigning impairment and some who were not, some credible cases were inevitably assigned (incorrectly) to the non-credible group. This bias may artificially lower the threshold at which cases would be considered non-credible. Third, Victor et al removed any cases suspected of having low IQ or dementia. This exclusion criterion also reduces the observed FPR in the credible group.

Finally, Larrabee cites Schroeder and Marshall (2011) as showing a lower FPR than predicted by Berthelson et al (2013). Schroeder and Marshall (2011), however, excluded 38% of their original sample due to low IQ (defined as IQ < 80), thus excluding patients more likely to fail PVTs. Supporting that interpretation, Table 4 in the Schroeder and Marshall paper shows that the FPR was indeed unacceptable (14%) in the group with IQ between 80 and 89, and presumably the FPR would have been even higher in groups with even lower ability (those with IQ below 80). This highlights the fact that the actual FPRs for patients with no incentive vary widely depending on a diversity of factors, among which true ability level is important.

What can be concluded from these comparisons? Both the Davis and Millis (2014) and Larrabee (2014) papers correctly highlight that the binomial distribution may not offer an adequate description of the behavior of multiple PVTs, which are often highly skewed due to the performance of most people near “ceiling” on the individual PVTs. The form of these distributions creates additional instability in the selection of individual “cutoff” scores, because these values are selected from the tails of distributions where precision of measurement is lower than it is in the more central regions of the distribution. There is further imprecision in assuming uniform PVT inter-correlations, because the criteria are usually based on *any combination* of failures, but there is higher likelihood of joint failure on some tests, and lower likelihood of joint failure on others. An adequate model would therefore need to consider different criteria based on specific pairs, triads or higher-number groupings of failures on specific tests. Overall, Berthelson and colleagues’ simulation results are valuable in highlighting the possible impact of inter-correlation among PVTs, which increases the likelihood of false positive errors beyond those predicted under the conventional binomial theory. Unfortunately, however, these simulations probably fail to make accurate predictions not only due to deviations of the model from the empirical data in skewness and correlational structure (as noted above), but also in other ways. For example, the binomial models do not account for differences among diverse instruments in their psychometric properties (e.g., reliability, differential sensitivity at different levels of ability, kurtosis, and more), or the ways in which instruments may behave differently in different samples.

Negative Binomial Regression Model

As an alternative to the standard binomial, Davis and Millis (2014) use a negative binomial distribution as the basis for their regression model. This is a good choice given that the negative binomial generally offers a better fit to over-dispersed distributions (where variance is greater than the absolute value of the mean, as is often found in PVT data). But there remains no substantive evidence that the negative binomial provides a valid basis for interpreting data across diverse, heterogeneous collections of PVTs, or that it is superior to the standard binomial. We describe below the details of the analysis and interpretations in the Davis and Millis paper, to determine how it informs the arguments about setting criteria for overall FPRs based on multiple PVTs.

First, it should be recognized that the Davis and Millis paper failed to reject the null hypothesis of no relationship between FPR and number of PVTs, but this does not mean that the true effect of PVTs on failure rate is *equivalent* to zero. It may simply be that the statistical power was inadequate to detect the observed effect size. Indeed, given the restricted range of PVTs administered in this study, along with the low frequency and skewed distribution of PVT failures, statistical power for this analysis was necessarily low. To demonstrate that the true effect is inconsequentially different from zero would demand a much larger sample (Tryon & Lewis, 2008). So while this paper did fail to find sufficient evidence to claim a statistically significant effect, it should not be interpreted as proving there is no influence of number of PVTs on the FPR (i.e., “accepting the null hypothesis”). Instead, when one has failed to reject the null hypothesis, it is important to consider the observed effect sizes and determine whether these might be meaningful if they were in fact real. Particularly given the stakes involved in forensic neuropsychology exams, the conservative approach is to assure that a meaningful effect is not missed (that is, to avoid Type II errors).

This brings us to the second, more important point: *the effect size identified by* Davis and Millis (2014) *for number of PVTs is actually large, but it is interpreted as if it were not important, and as if it were smaller than other effects*. The paper states:

> "Considered in relation to the other predictors, the number of PVTs administered appears to be of less importance than the evaluation context and other participant characteristics” (Davis & Millis, 2014, page 208).

The index of effect size is the incidence risk ratio (IRR). The key comparison is that between the IRR of 1.2 for PVTs administered and the IRR of 1.6 for medicolegal context. The claim that the medicolegal context effect is larger is based on direct comparison of these IRRs (i.e., that 1.6 is greater than 1.2; and for each unit increase in the former there is 60% increase in risk, while for the latter the increase is only 20%). But a direct comparison of these parameter estimates is misleading, because medicolegal context is a *dichotomous* variable, so it can only increase by one unit. In contrast, the number of PVTs is a *continuous* variable, and there is no theoretical limit to the number of PVTs that can be administered. In forensic practice it is not unprecedented to see 10 or more PVTs administered. Assuming that the magnitude of this effect is close to that estimated in the Davis and Millis paper, then giving 3 extra PVTs will have a greater effect than medicolegal context on PVT failure (IRR

= 1.2^3 = 1.73 for 3 extra PVTs, compared to IRR = 1.6 for medicolegal context). If 6 extra PVTs are given, this will increase PVT failure rates almost three-fold (IRR = 2.98), approximately double the observed effect of medicolegal context. These results in the Davis and Millis paper are at odds with the statement that “the impact of the number of PVTs administered on the number failed would likely remain minimal” (page 208). In fact, the results reported in Davis and Millis (2014) indicate that increasing the number of PVTs can have a major impact on the number of failed PVTs. Again, it seems more likely that the lack of significance in this analysis was due to limited statistical power rather than the absence of a true effect.

In addition, the Davis and Millis paper states: "Consistent with previous research, contextual factors (i.e., clinical versus medico-legal) … demonstrated significant associations with the number of PVTs failed" (page 208). This statement appears despite the fact that medico-legal status was not a statistically significant predictor of failed PVTs (p = .073), and in this respect, medico-legal status was not any more influential than the number of PVTs administered. This is particularly interesting given the widespread assumption that a medico-legal context is the *most* important predictor of the motivation to perform poorly and to fail PVTs. Applying the same logic that the Davis and Millis paper used to suggest that number of PVTs administered does not affect PVT failure rate (that it is not a significant predictor), should we similarly conclude that medico-legal context has no impact on PVT failure? That would contradict the primary assumption underlying most published research on PVTs. Instead, the data suggest that medico-legal status, the number of PVTs, and other ability factors are all important, but the Davis and Millis study lacked statistical power to declare significance for all these effects.

Summary and Conclusions

In summary, we believe the recent papers about how to interpret the results of multiple PVTs raise very important concerns, and indicate that reliance on references to theoretical models, or on any uniform set of criteria is premature, and not supported by existing evidence. Some major issues include:

1. *We lack adequate theoretical, mathematical, or statistical models of PVT failure across multiple tests, so we need to rely on empirical data*: So far most of the work on the likelihood of PVT failure is based on variants of binomial models, which appear to offer a poor fit to empirical data. Berthelson et al. (2013) make a valid point that the FPR is increased by inter-correlations among tests. Larrabee (2014) offers cogent criticisms about why Berthelson’s revised model is a poor fit for real data on PVTs (given the over-dispersion of these data), but no one has presented an alternative theoretical model that is superior. Even if we accept the Davis and Millis (2014) model, we face the same concerns about impact of number of tests that were posed by Berthelson and colleagues, and add additional concerns about education level and disability status. Until and unless we develop superior models, practice must be dictated by empirical data. Unfortunately the empirical data so far are incomplete, fail to cover many real-world scenarios, and therefore need to be

considered very carefully before application to individual cases with high-stakes outcomes.

2. *The empirical data show a clear impact of number of PVTs administered*: All available evidence indicates that the number of PVTs makes a substantial difference in the overall FPR. The empirical data recently presented in the Davis and Millis (2014) and Larrabee (2014) papers suggest that with 7 or more PVTs the criterion of 2 or more failed PVTs is not stringent enough to keep false positives at an acceptable level (e.g., FPR < 10%). When administering 8 or more PVTs within a single battery, application of a 2-failed-PVT criterion is not supported by any of these theoretical models or empirical data.

3. *The empirical data show a clear impact of patient factors, including general ability*: The threats to generalization from empirical studies include well-demonstrated general ability affects (even within “normal” ranges; see both Davis & Millis, 2014; and Schroeder & Marshall, 2011). The range of linguistic, cultural, or situational characteristics that may further affect interpretation are just beginning to be explored, but already indicate that a high degree of caution should be applied before interpreting PVT failures in individual cases that diverge on any patient characteristics from a given reference group. Existing data, including the findings of Davis and Millis (2014), refute claims that PVTs have no relation to other tests of neuropsychological abilities.

4. *The existing empirical study designs must be examined carefully before applying specific rules to an individual case*: The overall FPR may vary markedly across empirical studies given differences in: (a) inclusion/exclusion criteria; (b) combinations of tests used (which in turn may possess different correlations with each other and different dispersions of scores within each test); and (c) criteria used for each individual test. For example, the Schroeder and Marshall sample excluded anyone with IQ below 80, and showed unacceptably high FPR in the IQ range of 80 to 89. Thus, although some may be inclined to cite this study as evidence that using a 2-failed PVT rule is appropriate even in psychotic patients, this would be inappropriate in the majority of people with chronic schizophrenia, who according to at least one meta-analysis, have an average IQ of approximately 83.5 (Heinrichs & Zakzanis, 1998).

5. *Empirical demonstrations should be evaluated carefully for redundancy between criteria for case selection and the methods for calculating false positive rates*: A common research design in the study of performance validity has been the “known groups design,” which defines “credible” groups in part based on PVT performance (e.g., by including only patients who fail fewer than 2 PVTs, following the Slick criteria for “probable malingering”; Slick, Sherman, and Iverson, 1999), and then determines the FPR based on how many PVTs are failed. For example, as discussed earlier, in the Victor et al (2009) paper, the inclusion/exclusion criteria were based on freestanding PVTs while the FPR calculations used embedded PVTs. To the extent that freestanding and embedded PVTs are measuring the same construct this is an experimental design flaw and introduces systematic bias. At best this practice

> demonstrates the concurrent validity of two different measures of the same construct. At worst it reflects circular logic and a direct confounding of inclusion-exclusion criteria with outcome variables (sometimes referred to as "criterion contamination"). It will benefit the discipline to assess the broader literature on PVTs beyond the examples on which we focus here, and assess the potential impact of these design issues on current practice.

The articles by Berthelson et al., Davis and Millis, and Larrabee have brought much needed attention to the problems of interpreting multiple PVTs in clinical neuropsychology. We believe that simulations of PVT failure rates are useful to illustrate the possible effects of diverse properties of neuropsychological batteries, examinee characteristics, and contextual factors that deserve attention. Unfortunately, the models do not yet provide a reliable basis for setting practical criteria for validity, and we therefore need to base these decisions directly on empirical data. The existing empirical data, however, do not support generalized application of any criteria that fail to consider the number of tests, the specific tests used in combination, the population in which the validation was conducted, and the design of the validation study. We hope that future practice will strive to avoid applying criteria to individual cases, when these criteria are based on inappropriate generalizations from either theoretical or empirical observations.

Acknowledgments

Dr. Bilder has received honoraria and/or consulting fees within the last two years from EnVivo, Johnson & Johnson, Lundbeck, Takeda, and Think Now, Inc.; he is a shareholder/advisor to bLife, Inc. Over the last 10 years Dr. Bilder has received fees as an expert in forensic neuropsychology: he has given deposition and/or trial testimony in cases for which he was retained by attorneys for plaintiff (46% of cases) and defense (54% of cases).

The preparation of this manuscript was supported by the Michael E. Tennenbaum Family Center for the Biology of Creativity and a grant from the National Institute of Mental Health (R01MH101478). We thank Erin Bigler, Lori Holt, Justin B. Miller, and Jeffrey A. Schaeffer for valuable comments on earlier versions of this manuscript.

References

Berthelson, Lena; Mulchan, Siddika S.; Odland, Anthony P.; Miller, Lori J.; Mittenberg, Wiley. False positive diagnosis of malingering due to the use of multiple effort tests. Brain Injury. 2013; (0):1–8. [PubMed: 23252433]

Boone KB, Lu P. Noncredible cognitive performance in the context of severe brain injury. Clin Neuropsychol. 2003; 17(2):244–254. [PubMed: 13680432]

Davis JJ, Millis SR. Examination of Performance Validity Test Failure in Relation to Number of Tests Administered. Clin Neuropsychol. 2014; 28(2):199–214. [PubMed: 24528190]

Heinrichs RW, Zakzanis KK. Neurocognitive deficit in schizophrenia: A quantitative review of the evidence. Neuropsychology. 1998; 12(3):426–445. [PubMed: 9673998]

Larrabee GJ. Detection of malingering using atypical performance patterns on standard neuropsychological tests. Clin Neuropsychol. 2003; 17(3):410–425. [PubMed: 14704892]

Larrabee GJ. Aggregation across multiple indicators improves the detection of malingering: relationship to likelihood ratios. Clin Neuropsychol. 2008; 22(4):666–679. [PubMed: 17886147]

Larrabee, Glenn J. Malingering scales for the Continuous Recognition Memory Test and the Continuous Visual Memory Test. The Clinical Neuropsychologist. 2009; 23(1):167–180. [PubMed: 18609339]

Larrabee, Glenn J. False-Positive Rates Associated with the Use of Multiple Performance and Symptom Validity Tests. Archives of Clinical Neuropsychology. 2014 Jun 29.(4):364–373. 2014 Epub 2014 Apr 24. [PubMed: 24769887]

Pearson. Advanced clinical solutions for WAIS-IV and WMS-IV: Administration and scoring manual. The Psychological Corporation, San Antonio. 2009

Pella, Russell D.; Hill, Benjamin D.; Shelton, Jill Talley; Elliott, Emily; Gouvier, Wm Drew. Evaluation of embedded malingering indices in a non-litigating clinical sample using control, clinical, and derived groups. Archives of Clinical Neuropsychology. 2012; 27(1):45–57. Epub 2011 Nov 9. [PubMed: 22075576]

Schroeder, Ryan W.; Marshall, Paul S. Evaluation of the appropriateness of multiple symptom validity indices in psychotic and non-psychotic psychiatric populations. The Clinical Neuropsychologist. 2011; 25(3):437–453. Epub 2011 Mar 2. [PubMed: 21391153]

Slick DJ, Sherman EM, Iverson GL. Diagnostic criteria for malingered neurocognitive dysfunction: proposed standards for clinical practice and research. Clin Neuropsychol. 1999; 13(4):545–561. [PubMed: 10806468]

Tryon, Warren W.; Lewis, Charles. An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. Psychological Methods. 2008; 13(3): 272. [PubMed: 18778155]

Victor TL, Boone KB, Serpa JG, Buehler J, Ziegler EA. Interpreting the meaning of multiple symptom validity test failure. Clin Neuropsychol. 2009; 23(2):297–313. doi:903042154 [pii] 10.1080/13854040802232682. [PubMed: 18821138]

Table 1

Analysis of Data from Davis and Millis, Compared to Berthelson Simulation Results for 6, 7, and 8 PVTs Given

	Number of PVT Failures			Compare 0 vs 1 or more failed		Compare 0 and 1 vs 2 or more failed	
Counts	0	1 or more	2 or more	χ^2	p	χ^2	p
Observed	53	26	9	3.4	0.07	0.00	1.00
Expected	60	19	9				
Percent				Z	p	Z	p
Observed	67	33	11	1.87	0.06	0.17	0.87
Expected	76	24	12				

Note. Observed values computed from Table 5 of Davis and Millis (2014); expected values computed from Table IV of Berthelson and colleagues (2013). Counts and percent data do not yield identical statistics due to rounding errors.

Table 2

Overall False Positive Rates (%) from Theoretical Models and Empirical Data as a Function of the Number of PVTs Given and Criteria for Judging Invalidity

	Criterion: Number of failed PVTs to judge battery invalid					
Number of PVTs Given **Source**	**>=1**	**>=2**	**>=3**	**>=4**	**>=5**	**>=6**
5 PVTs						
Ind Binom	*41.0*	*8.2*	*0.9*	*0*	*0*	*0*
Corr Binom	*33.8*	*11.5*	*3.6*	*0.9*	*0.2*	*0*
Neg Binom-14	*26.7*	*7.1*	*1.9*	*0.5*	*0.1*	*0.0*
Neg Binom-12	*33.8*	*11.4*	*3.8*	*1.3*	*0.4*	*0.1*
ACS-Data	25.0	6.0	1.0	0	0	0
DM-Data	40.0	40.0	0	0	0	0
Larrabee-Data	41.6	5.6	0	0	0	0
6 PVTs						
Ind Binom	*46.9*	*11.4*	*1.6*	*0.1*	*0*	*0*
Corr Binom	*33.8*	*11.5*	*3.6*	*0.9*	*0.2*	*0*
Neg Binom-14	*30.4*	*9.3*	*2.8*	*0.9*	*0.3*	*0.1*
Neg Binom-12	*38.0*	*14.4*	*5.5*	*2.1*	*0.8*	*0.3*
DM-Data	26.9	3.9	0	0	0	0
7 PVTs						
Ind Binom	*52.2*	*15.0*	*2.6*	*0.3*	*0*	*0*
Corr Binom	*41.1*	*17.5*	*7.3*	*2.9*	*1*	*0.3*
Neg Binom-14	*34.4*	*11.9*	*4.1*	*1.4*	*0.5*	*0.2*
Neg Binom-12	*42.4*	*17.9*	*7.6*	*3.2*	*1.4*	*0.6*
DM-Data	33.3	15.2	3.0	0	0	0
Larrabee-Data	48.0	11.0	4.0	0	0	0
8 PVTs						
Ind Binom	*57.0*	*18.7*	*3.8*	*0.5*	*0*	*0*
Corr Binom	*44.0*	*20.2*	*9.3*	*4.1*	*1.7*	*0.6*
Neg Binom-14	*38.7*	*15.0*	*5.8*	*2.2*	*0.9*	*0.3*

	Criterion:					
	Number of failed PVTs to judge battery invalid					
Number of PVTs Given **Source**	**>=1**	**>=2**	**>=3**	**>=4**	**>=5**	**>=6**
Neg Binom-12	*46.9*	*22.0*	*10.3*	*4.8*	*2.3*	*1.1*
DM-Data	40.0	15.0	15.0	15.0	15.0	5.0

Note. Estimates based on theoretical models are shown in italics; empirical data are underlined. Theoretical models include: the independent binomial (Ind Binom); correlated binomial (Corr Binom); and the negative binomial model of Davis and Millis (2014)(this model was estimated with parameters as follows: age = 43, ADL = 0.7, and medicolegal status = 0 [non-medicolegal], at two different values for education: 14 years of education (Neg Binom-14), and 12 years of education (Neg Binom-12). Empirical data include the Pearson ACS data (ACS-Data), the empirical data from Davis and Millis (DM-Data), and the Larrabee data (Larrabee-Data). DM-Data for 5 PVTs are based on only 5 cases. FPR Values exceeding 10% are in darker shading; FPR values greater than 5% but less than 10% are in lighter shading.