Neuropsychology Review (2023) 33:581–603
https://doi.org/10.1007/s11065-023-09601-7

REVIEW

Review of Statistical and Methodological Issues in the Forensic Prediction of Malingering from Validity Tests: Part I: Statistical Issues

Christoph Leonhard[1]

Received: 24 April 2021 / Accepted: 29 March 2023 / Published online: 24 August 2023

Abstract

Forensic neuropsychological examinations with determination of malingering have tremendous social, legal, and economic consequences. Thousands of studies have been published aimed at developing and validating methods to diagnose malingering in forensic settings, based largely on approximately 50 validity tests, including embedded and stand-alone performance validity tests. This is the first part of a two-part review. Part I explores three statistical issues related to the validation of validity tests as predictors of malingering, including (a) the need to report a complete set of classification accuracy statistics, (b) how to detect and handle collinearity among validity tests, and (c) how to assess the classification accuracy of algorithms for aggregating information from multiple validity tests. In the Part II companion paper, three closely related research methodological issues will be examined. Statistical issues are explored through conceptual analysis, statistical simulations, and through reanalysis of findings from prior validation studies. Findings suggest extant neuropsychological validity tests are collinear and contribute redundant information to the prediction of malingering among forensic examinees. Findings further suggest that existing diagnostic algorithms may miss diagnostic accuracy targets under most realistic conditions. The review makes several recommendations to address these concerns, including (a) reporting of full confusion table statistics with 95% confidence intervals in diagnostic trials, (b) the use of logistic regression, and (c) adoption of the consensus model on the "transparent reporting of multivariate prediction models for individual prognosis or diagnosis" (TRIPOD) in the malingering literature.

Keywords Performance validity tests · Measurement · Test accuracy · Test validity · Bias · Malingering

Introduction

Past decades have seen a steady increase in the scholarly interest in malingering, with about 713 related APA Psychinfo® entries during the 1990s, rising to 1625 during the 2010s. By 2015, about 25% of publications in neuropsychology journals were devoted to malingering (Martin et al., 2015). Interest in the forensic determination of malingering is commensurate with its tremendous social, legal, and economic importance. A positive finding of malingering in a neuropsychological exam may, for example, lead to an examinee being denied a large financial settlement in a tort case or receiving a harsher criminal sentence including the death penalty (cf. Myers et al., 2016). Economically, the cost of malingering related to disability cases in the US has been estimated to exceed $20 billion in 2011 alone (Chafetz & Underhill, 2013). However, an econometric analysis found that the cost of providing disability payments to claimants who are malingering was only slightly greater than the savings from mistakenly rejecting claimants who are not malingering (Benitez-Silva et al., 2004, p. 25).

When identifying the presence of malingering of neuropsychological dysfunction, somatic symptoms, psychiatric presentation, or mixed symptom presentation in examinees undergoing forensic[1] neuropsychological evaluations, the most

✉ Christoph Leonhard
chleonhard@yahoo.com

[1] The Chicago School of Professional Psychology at Xavier University of Louisiana, Box 200, 1 Drexel Dr, New Orleans, LA 70125, USA

[1] The term "forensic" shall refer to any situation where there is an actual or potential legal question regarding the veracity of an examinee´s presentation such as when making determinations about the presence of sexual abuse, eligibility for disability benefits, fitness for fulfilling a particular role, such as being a parent or police officer, ability to stand trial, presence of neurocognitive and/or psychiatric conditions that may affect guilt, innocence, or sentencing of a criminal defendant, tort situations involving questions of the veracity of

recent neuropsychology assessment guidelines (Sherman et al., 2020), while not universally accepted by forensic neuropsychologists (Sweet et al., 2021, p. 1071), describe a diagnostic algorithm that calls for multidimensional criteria to be evaluated in four categories, including (a) the presence of an external incentive, (b) invalid presentation on an examination that is indicative of feigning or exaggeration, (c) marked discrepancies between obtained test data or symptom reports and other types of evidence, and (d) an inability to fully account for behaviors that meet the invalid presentation criteria by another developmental, medical, or psychiatric condition (Sherman et al., 2020, Box 1). This algorithm is an update of those previously published by the same group (Slick & Sherman, 2012; Slick et al., 1999). The 1999 version of the Slick et al. algorithm has been widely adopted (Sherman et al., 2020) and was considered by the American Academy of Clinical Neuropsychology (AACN) to be more representative of the state of neuropsychological knowledge than criteria for the determination of malingering of the American Psychiatric Association (2000) under DSM-IV-TR (Heilbronner et al., 2009, p. 1098).

While originally developed to detect malingering of neurocognitive dysfunction, malingering determination algorithms, and the statistical and research methods involved in their validation and use in forensic settings, have since heavily impacted neighboring areas in assessment, such as methods to ascertain the veracity of reports of child sexual abuse (Proeve, 2009), feigning of somatic (Bianchini et al., 2005) and psychiatric complaints (Edens et al., 2020; Sherman et al., 2020), and diagnosis of factitious disorder (Chafetz et al., 2020). However, the scope of both parts of this review is expressly limited to the validity of performance validity tests (PVTs) to determine malingered neurocognitive dysfunction in forensic neuropsychological exams and eschews discussion of the use of validity tests for other purposes, such as in the context of clinical neuropsychological exams where there are no external incentives on examinees to present a certain way.

This two-part review focuses on statistical and methodological issues in the forensic determination of malingering from knowledge of PVT scores. Depending on the type of malingering being determined, these could be free standing PVTs, such as the Test of Memory Malingering (TOMM; Tombaugh, 1996) or embedded PVTs, such as Reliable Digit Span (RDS; Greiffenstein et al., 1994). While, conceptually, all validity tests function nearly identically in the determination of malingering, this paper focusses primarily on PVTs, except when prior research discusses PVTs and symptom validity tests without distinguishing among them. Well over 50 freestanding and embedded validity tests have been developed (see Sweet, 2009, pp. 585–608). Consideration of validity test scores as part of the forensic determination of malingering appears in criteria B1, B2, and C5 in the 1999 and 2012 malingering determination algorithms (Slick & Sherman, 2012; Slick et al., 1999), and in criteria B1b, B1c, B2b, B3b, B4b, and B4c in the 2020 algorithm (Sherman et al., 2020). One survey found that 92% of neuropsychologists report "often" or "always" using validity tests (Martin et al., 2015), and another found that 74% of neuropsychologists believe they are "often" or "always" able to recognize malingering (Aita et al., 2020). Among forensic neuropsychologists, 99% consider use of validity tests mandatory (Martin et al., 2015, p. 748).

In recent years, use of validity tests has expanded beyond forensic settings. Professional bodies such as the National Academy of Neuropsychology (NAN) and the American Academy of Clinical Neuropsychology (AACN) call for the use of validity tests in all assessment settings. NAN, for example, advises that "adequate assessment of response validity is essential in order to maximize confidence both in the results of ability measures and in the diagnoses and recommendations that are based on the results" (Bush et al., 2005, p. 425). Similarly, in its Consensus Conference Statements, AACN recommends the use of validity measures in all evaluations (Heilbronner et al., 2009, p. 1121; Sweet et al., 2021, p. 1066). Finally, while not specifically recommending use of PVTs, the American Medical Association's (AMA) *Guides to the Evaluation of Permanent Impairment* advise examiners to always be aware of the possibility of malingering when evaluating impairments secondary to mental or behavioral disorders (Rondinelli et al., 2008, p. 353).

Throughout both parts of the review, the following naming conventions are used. The presence of malingering is denoted M^{+}, and the absence of malingering is denoted M^{-}. A positive finding on a PVT indicative of M^{+} is denoted PVT^{+}, while a negative finding on a PVT indicative of M^{-} is denoted PVT^{-}. Note that in prior writing (e.g., Chafetz, 2011; Larrabee, 2008) the wording "failure on" or "failing" a PVT is sometimes used to denote a positive finding on a PVT (PVT^{+}). "Passing a PVT" is sometimes used to denote a negative finding on a PVT (PVT^{-}) indicative of credible responding (M^{-}). In keeping with the terminology from the medical diagnostics literature that a positive result on a test indicates presence of the attribute tested, this paper will use the term "positive finding on a PVT" (PVT^{+}) to indicate "failing" a PVT. "Negative finding on a PVT" (PVT^{-}), will indicate "passing" the PVT.

Note that these dichotomizations are almost certainly an oversimplification of a range of underlying presentations

Footnote 1 (continued)

neurocognitive abilities and related conditions. Such actual or potential legal questions arise most frequently in medicolegal settings, but may arise also in clinical contexts (cf. Sherman et al., 2020, p. 9; Sweet et al., 2021, p. 1059).

among forensic neuropsychological examinees. At one end of the continuum, for example, an examinee with a mild traumatic brain injury (mTBI) who is a college athlete might deny symptoms of neurocognitive dysfunction to avoid putting their position on the team in jeopardy. Next on the continuum might be a completely honestly responding examinee. Next, an examinee who is troubled by mild neurocognitive symptoms following mTBI but somewhat embellishes these symptoms to ensure concerning symptoms are not missed. Next on the continuum might be an examinee who grossly exaggerates mild neurocognitive symptoms, and at the other extreme of the continuum, an examinee with no or minimal symptoms who purposefully malingers severe dysfunction.

These naming conventions also raise the conceptual question of whether a finding of PVT^{-} implies M^{-}, or merely that no conclusion can be drawn regarding an examinee's malingering status (see Chafetz et al., 2020 for related discussion). Because computation of classification accuracy statistics widely used in the biomedical diagnostics and malingering literatures, such as sensitivity and specificity, require that tests either rule in or rule out conditions, the working assumption in this paper is that PVT^{-} implies M^{-} and PVT^{+} implies M^{+}. Note that this does not imply that a PVT is required to rule in malingering with as much specificity as it possesses sensitivity to rule it out. Because of the inverse relationship between sensitivity and specificity, validity tests are designed to favor specificity over sensitivity (see e.g., Sweet et al., 2021, p. 1089). Parenthetically noted, only the best tests, those that are considered reference standards or gold standards, achieve both near perfect (reference standard) or perfect (gold standard) sensitivity and specificity. Note also that a determination of M^{-} does not necessarily mean an examinee is responding truthfully, it merely means the respondent is not determined to be malingering.

Analytic Strategy for this Review

In Part I, three statistical issues relevant to the prediction of malingering based on knowledge of PVTs are explored, including (a) the need for full confusion[2] table statistics and confidence intervals when reporting malingering classification accuracy statistics, (b) the quantification of information overlap among PVTs in determining malingering status based on scores from two or more PVTs, and (c) the expectations for classification accuracy of PVT aspects of malingering determination algorithms (Larrabee et al., 2007; Sherman et al., 2020; Slick & Sherman, 2012; Slick et al., 1999).

While this discussion is mostly theoretical and conceptual, specific examples from PVT validation studies are interspersed throughout. These studies were selected in two ways.

1. An exhaustive list of 60[3] validation studies from the first 20 or more years of validation literature on the TOMM was examined (see Part II of this review, Table 1 for complete list). Studies on the TOMM were chosen to exemplify statistical and research methodological issues in the malingering literature because the TOMM is by far the most commonly used PVT at 78% reported use among neuropsychologists (Martin et al., 2015, p. 762). The TOMM is also recognized as having a "…large and diverse research base…" (Martin et al., 2020, p. 88). Additionally, a reasonably up-to-date exhaustive list of TOMM validation studies was available (Martin, et al., 2020). Martin et al. developed this list after reducing an initial set of 539 potentially applicable TOMM studies following a well-defined and well-documented selection process (Martin et al., 2020, pp. 90–94). Studies were published between 1997 and mid-2019. No attempt was made to identify TOMM studies published after that cutoff to allow optimal comparability with the findings of Martin et al. (2020), because the intent of this review was to use TOMM studies to exemplify statistical and research methodological concepts and not to present a meta-analysis on the TOMM. There are also two studies among the original list that are not discussed in this review, and one study not in the original list that is discussed. Instead of Ashendorf et al. (2003), Ashendorf et al. (2004) was included because the former study does not present TOMM results while the latter does. It was therefore assumed Ashendorf et al. (2003) had been mistakenly cited in Martin et al. (2020). Another study was excluded because, upon examination, data related to TOMM validation referenced in Martin et al. (2020, p. 99) could not be found (Greiffenstein et al., 2008).

2. To cross-check exemplification of statistical and methodological issues from TOMM studies, non-systematically selected validation studies on other PVTs were also examined. While the selection of these studies was not methodical, special care was taken to include studies on the Word Memory Test (WMT) and MSVT, because at 59% and 28% respectively, neuropsychologists report using these PVTs second and third most frequently (Martin et al., 2015, p. 762).

Specific statistical criteria for considering validity test scores in the determination of malingering appear in the 1999 malingering determination algorithm (Slick et al., 1999) only for Criterion B1, where forced-choice PVTs are taken

[2] A "confusion table" or "confusion matrix" is a 2×2 diagnostic classification table (see Figs. 1 and 2). It is a special case of the 2×2 contingency table often used in psychological research to show the association of two binary variables.

[3] In Martin et al. (2020, pp. 99–100; Table 5), 53 TOMM studies are listed with an additional two studies listed on p. 112 for a total of 55 studies. However, there is inconsistency in how studies from articles that report multiple studies are listed in Martin et al. (2020). For consistency, in Parts I and II of the present review, each separate study, whether reported in a publication with other studies or by itself, is counted separately.

to be indicative of malingering only if performance is below chance at $p < 0.05$. In a 2012 update (Slick & Sherman, 2012), a statistical requirement was added to Criterion B2 (p. 123) that when determining malingering from knowledge of one or more PVTs, the combined posterior probability that an examinee's performance was significantly below actual ability level, when considering all PVT scores together, would have to be "high (≥ 0.95)". However, this update appears not to have been widely adopted. For the 2020 update (Sherman et al., 2020), statistical requirements were changed again to specify that PVTs considered alone or in combination must now have a low false positive rate ($\leq 10\%$), which equates to a specificity of > .9. The 2020 criteria no longer include a specific posterior probability threshold, though the assumption appears to be that the required specificity of > .9 will yield a posterior probability of no less than 51% (p. 14). Additionally, all algorithms (Sherman et al., 2020; Slick & Sherman, 2012; Slick et al., 1999) agree in requiring two or more PVTs to have indicated PVT^+ to meet PVT related malingering criteria. All three versions also allow for the exception that PVT related malingering criteria can be met if one forced-choice PVT, such as the Portland Digit Recognition Test (PDRT; Binder, 1993; Binder & Willis, 1991), TOMM, MSVT, or WMT (P. W. Green et al., 1996; P. W. Green, 2003), is in the significantly below-chance range. However, while earlier versions of the algorithm (Slick & Sherman, 2012; Slick et al., 1999) stipulate a minimum significance level of $p \leq .05$ for forced-choice PVTs, the required significance level has been dropped in the most recent version (Sherman et al., 2020), in favor of "…empirically derived cutoffs…" (p. 11). In practice, this means that scores on forced choice PVTs are now meeting PVT criteria for M^+ at a much relaxed "significance" level of $p \leq .20$, one-sided (Binder et al., 2014).

Statistical Issue A: Need for Full Confusion Table Statistics and Confidence Intervals

A recent article (Lange & Lippa, 2017) called on neuropsychologists to never interpret sensitivity and specificity "… in isolation without consideration of other clinical utility measures," such as positive predictive power.[4] However, in the malingering literature, the awareness of the need to present the full set of confusion table classification accuracy statistics, also known as "test operating characteristics," is sporadic. The 2020 malingering determination algorithm (Sherman et al., 2020) appears to assume a "… low false positive rate (i.e., 0.10)…" (p. 6; equating to a specificity > 0.9) will result in a positive predictive power of at least 0.51 (p. 14). The algorithm also specifically eschews requiring specific cutoffs for classification accuracy statistics other than false positives because "… realistically, most clinicians do not have easy access to … sophisticated classification accuracy statistics…" (pp. 14–15).

However, while sensitivity and specificity are important qualities of a diagnostic test and are thus of considerable interest to an examiner selecting PVTs to aid in the prediction of malingering status, the posterior probability or the positive predictive power is the statistic that informs the examiner of the probability that a given examinee has been correctly determined to be M^+ given a finding of PVT^+ (Fletcher et al., 2014, pp. 117–118). This is true regardless of whether the determination of malingering status is based on just one PVT or on a group of PVTs considered together.

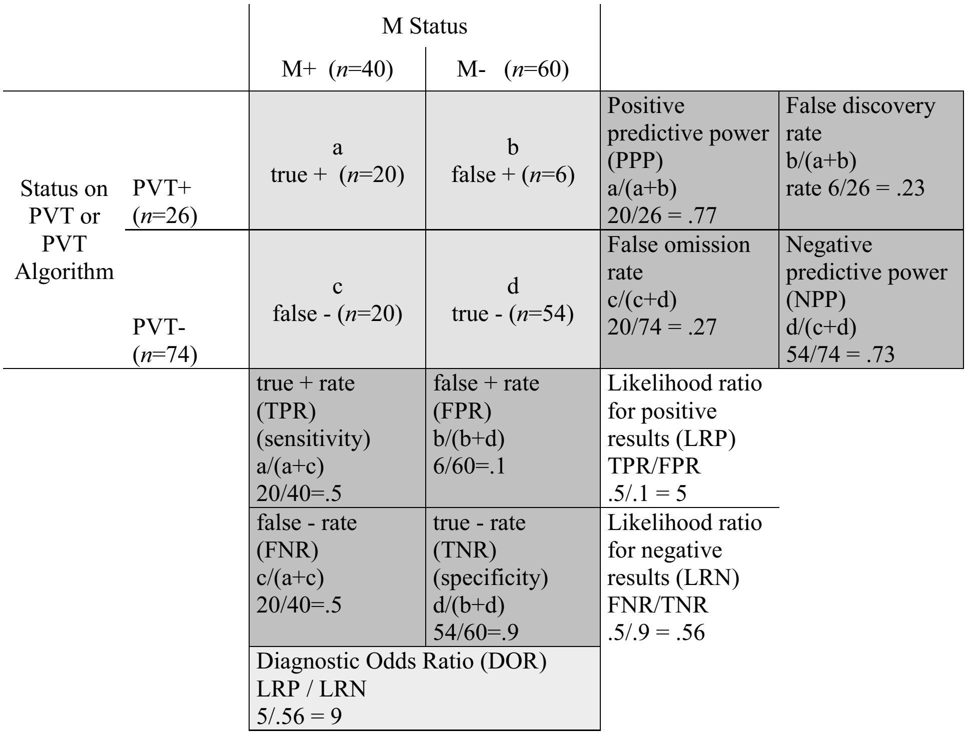
To examine this further, consider Figs. 1 and 2, which depict hypothetical confusion tables on the prediction of malingering status in a group of 100 neuropsychological examinees using a PVT or combination of PVTs with a sensitivity of 0.5 and a specificity of 0.9. Figure 1 assumes a base rate of malingering of 0.4, which is commonly reported in forensic settings (Larrabee, 2003, 2008; Larrabee et al., 2009; Sherman et al., 2020).[5] Figure 2 assumes a base rate of malingering of 0.1, the base rate of malingering considered common in clinical settings (Mittenberg et al., 2002).

Figures 1 and 2 illustrate many important classification accuracy statistics that can be calculated in a confusion table. Table calculations based on column values are identical across Figs. 1 and 2, while those calculated based on row values differ. Important column-based calculations include sensitivity, specificity, the likelihood ratio for positive results, the likelihood ratio for negative results, and diagnostic odds ratio (OR) (Glas et al., 2003). These calculations can be considered attributes of the predictor or predictors, in this case, of the PVT or of the PVT-based malingering determination algorithm. The most important row value is the positive predictive power, also known as posterior probability. This represents the likelihood that an examinee with a finding of PVT^+ is actually M^+.

[4] The terms "positive predictive power" and "posterior probability" are sometimes considered conceptually distinct in that positive predictive power is an operating characteristic of a test in a given setting, computed as the ratio of true positive cases over the total number of all examinees with a positive diagnostic finding. In comparison, posterior probability is the probability a given examinee with a positive test result is actually a true positive case (cf. Fletcher et al., 2014, p. 118). Despite this conceptual distinction, the computation of both classification accuracy statistics is identical. In this article, these terms will be used in keeping with their conceptual distinction.

[5] Knowledge of the base rate of a condition is required for the calculation of classification accuracy statistics. Statistical modelling and analyses of hypothetical simulation data in this review will be based on commonly accepted estimates of base rates from the malingering literature. Examination of whether these estimates are tenable considering the findings of the present review is beyond the scope of this review.

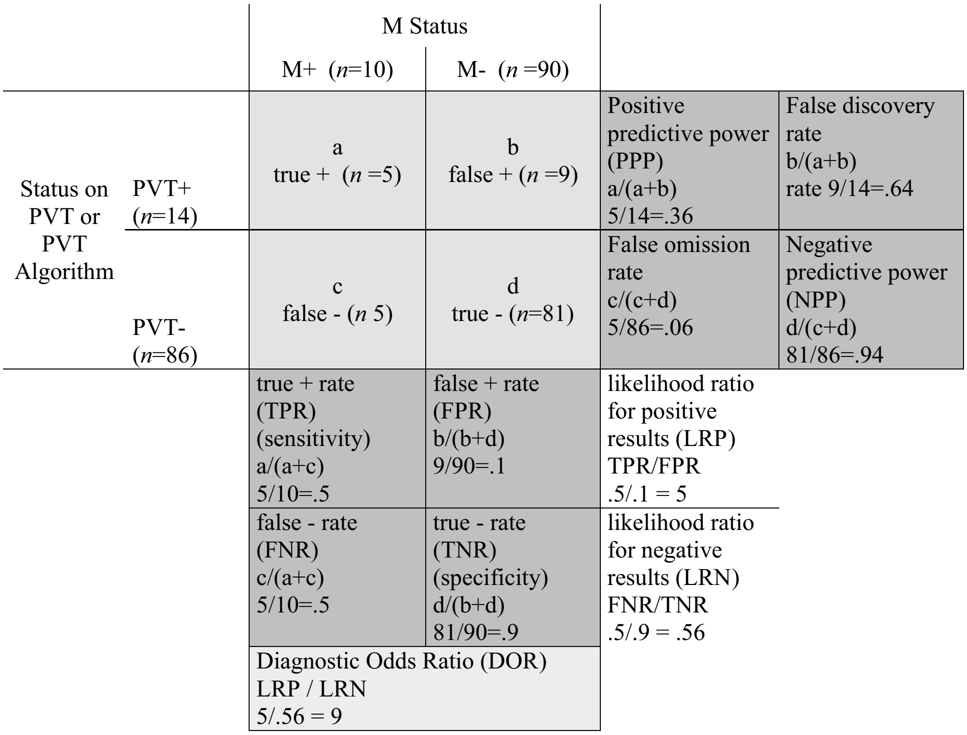
Fig. 1 Hypothetical confusion table for predicting malingering (M) from knowledge of a single performance validity test (PVT) or PVT algorithm with sensitivity = .5, specificity = .9, base rate of M+ = .4, and *N* = 100



		M Status			
		M+ (*n*=40)	M- (*n*=60)		
Status on PVT or PVT Algorithm	PVT+ (*n*=26)	a true + (*n*=20)	b false + (*n*=6)	Positive predictive power (PPP) a/(a+b) 20/26 = .77	False discovery rate b/(a+b) rate 6/26 = .23
	PVT- (*n*=74)	c false - (*n*=20)	d true - (*n*=54)	False omission rate c/(c+d) 20/74 = .27	Negative predictive power (NPP) d/(c+d) 54/74 = .73
		true + rate (TPR) (sensitivity) a/(a+c) 20/40=.5	false + rate (FPR) b/(b+d) 6/60=.1	Likelihood ratio for positive results (LRP) TPR/FPR .5/.1 = 5	
		false - rate (FNR) c/(a+c) 20/40=.5	true - rate (TNR) (specificity) d/(b+d) 54/60=.9	Likelihood ratio for negative results (LRN) FNR/TNR .5/.9 = .56	
		Diagnostic Odds Ratio (DOR) LRP / LRN 5/.56 = 9			

Note. "Confusion Table" = 2 x 2 contingency table to evaluate the accuracy of diagnostic classifications; M = Malingering; + = positive for; - = negative for; PVT = Performance Validity Test; PPP is also known as posterior probability.

Fig. 2 Hypothetical confusion table for predicting malingering (M) from knowledge of a single performance validity test (PVT) or PVT algorithm with sensitivity = .5, specificity = .9, base rate of M+ = .1, and *N* = 100



		M Status			
		M+ (*n*=10)	M- (*n* =90)		
Status on PVT or PVT Algorithm	PVT+ (*n*=14)	a true + (*n* =5)	b false + (*n* =9)	Positive predictive power (PPP) a/(a+b) 5/14=.36	False discovery rate b/(a+b) rate 9/14=.64
	PVT- (*n*=86)	c false - (*n* 5)	d true - (*n*=81)	False omission rate c/(c+d) 5/86=.06	Negative predictive power (NPP) d/(c+d) 81/86=.94
		true + rate (TPR) (sensitivity) a/(a+c) 5/10=.5	false + rate (FPR) b/(b+d) 9/90=.1	likelihood ratio for positive results (LRP) TPR/FPR .5/.1 = 5	
		false - rate (FNR) c/(a+c) 5/10=.5	true - rate (TNR) (specificity) d/(b+d) 81/90=.9	likelihood ratio for negative results (LRN) FNR/TNR .5/.9 = .56	
		Diagnostic Odds Ratio (DOR) LRP / LRN 5/.56 = 9			

Note. "Confusion Table" = 2 x 2 contingency table to evaluate the accuracy of diagnostic classifications; M = Malingering; + = positive for; - = negative for; PVT = Performance Validity Test; PPP is also known as posterior probability.

For a forensic neuropsychologist determining the malingering status of a particular examinee, the key statistic is arguably the posterior probability, also known as the positive predictive power, because it estimates the accuracy of the malingering determination in a particular examinee given a finding of PVT^+ (Fletcher et al., 2014, p. 118; Glas et al., 2003). It is beyond the scope of this paper to discuss whether it is acceptable to require a determination algorithm to have a positive predictive power of only 0.51 (cf. Sherman et al., 2020, p. 14), or in other words, to knowingly adopt a malingering determination algorithm that falsely labels 49% of PVT^+ examinees as M^+, a finding which may have serious negative consequences for the examinee. But, as Fig. 2 shows, if the only classification accuracy requirement is a specificity of 0.9, in clinical settings this leads to an expected positive predictive power of just 0.36. In other words, in clinical settings, the 2020 algorithm (Sherman et al., 2020) would accept a false discovery rate which falsely determines 64% of PVT^+ examinees to be M^+. Note that, because the 2020 algorithm (Sherman et al., 2020) does not specify a sensitivity requirement, the actual positive predictive power could be even lower. For example, based on the average sensitivity of 0.15 from simultaneous consideration of three PVTs reported by Chafetz (2011, p. 1247; Table 3), and the specificity of 0.9 required by the 2020 algorithm, in a clinical setting (base rate = 0.1), 86% of PVT^+ examinees would be falsely determined to be M^+ (positive predictive power = 0.14), while in a forensic setting (base rate = 0.4) a false determination of M^+ would be expected in 50% of PVT^+ examinees (positive predictive power = 0.50).

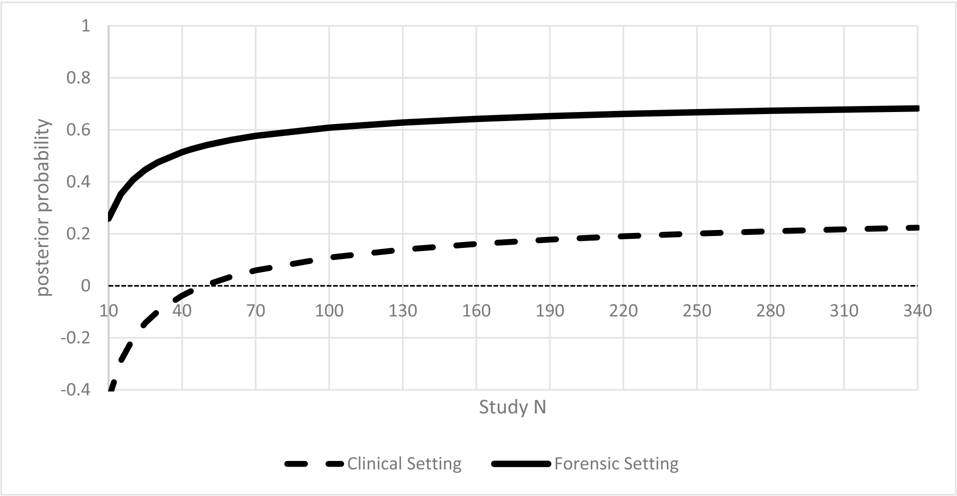
Therefore, algorithms for determining malingering (e.g., Larrabee et al., 2007; Sherman et al., 2020; Slick et al., 1999) may have unacceptably low positive predictive power when positive predictive power requirements are not specified, particularly in settings with a low base rate. This may occur despite the specific recommendation by NAN (Bush et al., 2005, p. 423) and AACN (Heilbronner et al., 2009, p. 1121; Sweet et al., 2021, p. 1076) to use PVTs to detect malingering in all settings, including clinical settings where the base rate of M^+ is low.

Another important statistic calculated from confusion tables is the diagnostic OR. It is obtained by dividing the likelihood ratio for positive results by likelihood ratio for negative results (Glas et al., 2003). A diagnostic determination algorithm or test with a diagnostic OR of 1 is "uninformative" and does not improve prediction over baseline estimation. If the diagnostic OR < 1, a positive diagnostic determination based on the algorithm or test lowers the odds of being positive for the condition compared the baseline estimation. A diagnostic OR > 1 means the determination algorithm or test is informative and improves prediction over the baseline estimation, whereby the odds of being positive for the condition increase if there is a positive finding on the test or testing algorithm. The diagnostic OR is also valuable for comparing two diagnostic determination algorithms against each other (Bossuyt et al., 2013, p. 11). Reporting the diagnostic OR is preferred over the use of the likelihood ratio for positive results (Glas et al., 2003; cf. Grimes & Schulz, 2005), because the likelihood ratio for positive results describes only the classification ability of positive results, and behaves erratically as specificity approaches 1. Because classification algorithms based on continuous data can always be made to have a specificity near 1 if cutoffs are set so that most cases are classified negative, the likelihood ratio for positive results can thus be seemingly impressively large even for poor classification algorithms if few cases are classified positive. The diagnostic OR, on the other hand, describes classification accuracy based on both positive and negative results. When there are empty cells in the confusion table that would make the diagnostic OR undefined due to division by zero, the value 0.5 is added to each cell count to estimate the diagnostic OR (Glas et al., 2003).

Classification accuracy statistics, like all statistics, are estimates of population parameters calculated from sample data. As such, they should be reported with 95% confidence intervals. (cf. Fischer et al., 2003; Glas et al., 2003). Sensitivity, specificity, the likelihood ratio for positive results, the likelihood ratio for negative results, and the positive predictive power are proportions (p), thus their 95% confidence interval can be calculated with the formula $p \pm 1.96\sqrt{p(1-p)/N}$ (Fletcher et al., 2014, p. 117; online calculators available, e.g., The Chinese University of Hong Kong, n.d.). The 95% confidence interval for the diagnostic OR is calculated with a more complex logarithmic formula (cf. Lawson, 2004), but can be readily obtained with free calculators online (e.g., Sci Stat, n.d.).

Among the more than 100 PVT validation studies examined for this review, not one study reported confidence intervals for classification accuracy estimates. To compute how large these unreported intervals would have been, sample calculations were made using hypothetical scenarios with the median PVT validation study sample size ($N = 44$) from Martin et al. (2020, Table 5). Assuming the typically reported specificity for PVTs of 0.9 (Sherman et al., 2020) for predicting malingering status from knowledge of PVT scores, the 95% confidence interval for specificity would be .9 ± .09, placing the lower bound below the 0.9 requirement from the 2020 malingering determination algorithm (Sherman et al., 2020). Confidence intervals for typically reported sensitivities are larger. Again assuming the typically reported sensitivity for PVTs of 0.5 (Sherman et al., 2020), the 95% confidence interval for sensitivity at $N = 44$ would be .5 ± .15. For PVT validation studies with small sample sizes, for example those with $N = 10$ (Rees et al., 1998; Shandera et al., 2010), the 95% confidence intervals would be .9 ± .19 for specificity and .5 ± .31 for sensitivity, again assuming typically reported specificities and sensitivities.

Fig. 3 Relationship of study N to the lower bound of the 95% confidence interval of the posterior probability of malingering (M^+) based on a single performance validity test (PVT) with sensitivity = .5 and specificity = .9 in forensic settings (base rate of $M^+ = .4$) and in clinical settings (base rate of $M^+ = .1$)



Note. Lower Bound of 95% Confidence Interval computed using the formula $posterior\ probability - 1.96\sqrt{p(1-p)/n}$ (Fletcher et al., 2014, p. 117) with $p = .77$ in forensic settings and $p = .36$ in clinical settings (see Figures 1 and 2 for derivation). Ratio of Study N to formula $n = N * .26$ in forensic setting and $n = N * .14$ in clinical settings (see Figures 1 and 2 for derivation).

The importance of reporting 95% confidence intervals is also evident when considering the posterior probability, or positive predictive power, which is the likelihood that an examinee with a finding of PVT^+ is indeed malingering. As can be seen in Fig. 1, in a forensic setting with a base rate of $M^+ = .4$, the point estimation of posterior probability is 0.77 given a finding of PVT^+. Figure 2 shows that, in a clinical setting with a base rate of $M^+ = .1$, the point estimation of posterior probability is 0.36 given a finding of PVT^+. Figure 3 shows the lower bound of the 95% confidence interval around these posterior probabilities in relation to study sample sizes.

As Fig. 3 shows, when the lower bound of 95% confidence intervals are calculated for PVT validation studies with small sample sizes, such as those with $N = 10$ (cf. Rees et al., 1998; Shandera et al., 2010), the posterior probability in clinical settings drops below zero, which means it is possible that examinees who score PVT^+ are actually *less* likely to be M^+ than those who score PVT^-. Also, regardless of sample size, the lower bound of the 95% percent confidence interval in clinical settings can never reach 0.51 (cf. Sherman et al., 2020, p. 14), where an examinee who scores PVT^+ would more likely than not be M^+. Assuming the mean sample size of all TOMM validation studies ($N = 44$; see above for derivation), when using the TOMM in forensic settings with the higher assumed base rate of $M^+ = .4$, the lower bound of the 95% confidence interval for the posterior probability given a finding of PVT^+ is about 0.53, assuming the test has a sensitivity of 0.5 and a specificity of 0.9. This means that, based on the findings of one of these studies with sample size $N = 44$, it is possible with $p > 0.05$ that such a PVT mistakenly classifies 47% of forensic examinees as M^+ when they are in fact M^-.

Calculation of the variability of classification accuracy statistics is also the basis for testing the statistical significance of the difference in determinative accuracy between different PVTs or between different malingering determination algorithms. Methods and formulas for doing so are readily available from statistical texts (e.g., Zhou et al., 2011, pp. 165–192). Here as well, none of the PVT validation studies examined for this article report statistical significance testing when comparing the determinative accuracy of PVT tests or variants of malingering determination algorithms.

In summary, when evaluating malingering determination algorithms and when publishing related test validation studies, presenting the full set of confusion table statistics is recommended. Furthermore, 95% confidence intervals should be reported with all classification accuracy estimations, and suspected differences in classification accuracy between PVTs or diagnostic determination algorithms should be tested for statistical significance. Doing so will present a complete picture of the ability of any given algorithm to aid in the determination of malingering, and will avoid giving the false impression that the accuracy of a malingering determination can be estimated with pin-point precision. Importantly, reporting this information will also allow practicing neuropsychologists to accurately estimate confidence limits of the posterior probability of malingering in any given forensic examinee.

Statistical Issue B: Estimating (Multi) collinearity Among PVTs in Determining Malingering Status Based on Scores from Two or More PVTs

To identify malingering, all algorithms require simultaneous positive findings from two or more validity tests, except if an examinee performs below chance on one forced-choice PVT (Larrabee et al., 2007; Sherman et al., 2020; Slick & Sherman, 2012; Slick et al., 1999). There has been discussion about the degree to which correlation among PVTs inflates the overall rate of false positives in the examination, limiting the incremental value of findings from additional PVTs beyond the first (e.g., Berthelson et al., 2013; Larrabee, 2014; Larrabee et al., 2019). Concerns about increased false-positives in evaluations due to overlapping information obtained from multiple PVTs are also reflected in the 2020 guidelines (Sherman et al., 2020), in which Criterion B1c requires "… taking into account the ratio of failed PVT scores to total number of PVTs administered" and also calls for "minimizing PVT redundancy" (p. 6). This is further quantified in the text that accompanies the guidelines, where the lower limit of the acceptable ratio of positive PVTs to PVTs administered is two PVTs with a positive finding to every seven PVTs administered (p. 14).

This concern about redundancy among PVTs as predictors of malingering resembles techniques used to identify multicollinearity and singularity among predictors in logistic regression (cf. Midi et al., 2010), linear regression (cf. Tabachnick et al., 2019, pp. 76–78), and in Bayesian regression (cf. Bayman & Dexter, 2021). Primarily, when predictors provide highly overlapping information across all these statistical prediction techniques, collinear predictors are either combined or eliminated, as neither frequentist nor Bayesian prediction can extract additional valuable information from redundant predictors.

Prior research on collinearity and the degree of information overlap among PVTs has shown mixed findings. High Pearson product-moment correlations have been reported among certain PVTs, especially if the PVTs are scored from the same items. For example, Digit Span scaled score and the RDS score have been found to correlate between $r = .83$ and $r = .92$ (Babikian et al., 2006, p. 152). However, average correlations of $r = .70$ were found even among seemingly unrelated validity tests (Bashem et al., 2014, p. 856), such as between TOMM Trial 2 and three other PVTs including Word Choice Test (WCT; Pearson, 2009), immediate recall on the Medical Symptom Validity Test, (MSVT; Green, 2004), and forced-choice hits in California Verbal Learning Test II (CVLT-II; Delis et al., 2000; Schwartz et al., 2016). In contrast, a recent study found that the average bivariate correlations among several PVTs was $r = .26$ (Larrabee et al., 2019), and a meta-analysis found that the average bivariate correlation among a group of approximately 30 PVTs was $r = .31$ (Berthelson et al., 2013). Other authors (e.g., Chafetz, 2011; Larrabee, 2008; Meyers et al., 2014) have asserted that PVTs, including TOMM, MSVT, RDS, and Meyers MMPI-2 Index should be considered *independent* predictors of malingering status, which would mean that their true correlation is not significantly different from $r = 0$ (cf. Baak et al., 2020, p. 2; Pepe, 2004, p. 197; Tabachnick et al., 2019, p. 7).

These contradictory findings are explored in the present review in three ways. First, related conceptual statistical issues are discussed. Second, a statistical simulation is conducted to see if test operating characteristics of typical PVTs allow for statistical independence. Third, data from previously published PVT validation studies are reanalyzed to explore the statistical independence of commonly used PVTs.

Conceptual and Statistical Issues Related to Independence of PVTs as Predictors of Malingering

When determining the degree of dependence or independence of PVTs in predicting malingering status, it is first necessary to define the universe of cases (U) where this information overlap is to be examined. In the malingering literature, the question of whether two or more PVTs contribute independent information to the prediction of malingering status is often examined by calculating Pearson product-moment correlations among pairs of PVTs in only the M^+ group, in only the M^- group, or in the M^+ group and the M^- group separately. For example, Larrabee (2008, p. 648, Table 5) calculated correlations among PVTs separately for the M^+ and M^- groups (the latter was labelled "TBI" in the original work), and Chafetz (2011, p. 1246, Table 2) and Jones (2013, p. 1050, Table 1) reported PVT associations for the M^+ group only. Such conditional independence calculations, however, fail to provide information about the independence of PVTs as predictors of malingering in the relevant U of cases, which includes all examinees irrespective of malingering status. To estimate this dependence or independence, the entire U of M^+ and M^- cases must be considered simultaneously.

To examine this issue further, consider Table 1, which depicts a hypothetical situation with 10 forensic examinees with the commonly reported forensic setting base rate for M^+ of 0.4. (cf. Larrabee, 2003, 2008; Larrabee et al., 2009; Sherman et al., 2020). There are also two hypothetical PVTs (PVT_1 and PVT_2) that predict malingering status perfectly if a cutoff for M^+ of $PVT \geq 10$ is used for both PVT_1 and PVT_2. The calculated Pearson r value for pairs of PVT_1 and PVT_2 scores only among M$^-$ examinees is $r = 0.00$.[6] For PVT_1 and PVT_2 pairs only among M$^+$ examinees the Pearson r value is

[6] This quantity may also be referred to as the conditional correlation of PVT_1 and PVT_2 conditioned on M$^-$.

Table 1 List of 10 hypothetical forensic examinees with known malingering status predicted from two validity measures

Examinee no	M status	PVT_1 score	PVT_2 score
1	M^-	1	0
2	M^-	1	1
3	M^-	1	0
4	M^-	0	0
5	M^-	0	0
6	M^-	0	1
7	M^+	11	10
8	M^+	11	11
9	M^+	10	10
10	M^+	10	11

M malingering, *PVT* performance validity test, PVT_1 first performance validity test, PVT_2 second performance validity test, M^- negative for M, M^+ positive for M

also $r = 0.00$,[7] which seemingly indicates that PVT_1 is independent of PVT_2 as a predictor of M status in a *U* consisting of both M^+ and M^- examinees. However, if the entire *U* of examinees is considered together when calculating the Pearson *r* value, the correlation of those same pairs of PVT_1 and PVT_2 scores is $r = 0.99$.[8] This finding demonstrates that PVT_1 and PVT_2 are completely collinear and thus fully redundant predictors of malingering status in *U,* which is the exact opposite of independent predictors. In other words, PVT_2 adds no information not already available from PVT_1 when predicting malingering status in *U*. This finding also proves that conditional correlations of PVT_1 and PVT_2 are not informative as to whether PVT_1 and PVT_2 provide additional valuable information in the prediction of malingering status.

While Table 1 depicts a hypothetical situation, it does demonstrate that evaluating the conditional independence of pairs of PVTs in homogeneous malingering-status groups is not a valid method to examine collinearity or information overlap among PVTs as predictors of malingering status.

In light of this finding, studies that contributed bivariate PVT correlations to the prior meta-analysis on this topic were reexamined for malingering status (Berthelson et al., 2013). As stated by Berthelson et al. (2013, p. 910), all selected studies included only M^- cases, namely "… participants that were not seeking compensation or involved in litigation." All correlations were calculated using data from homogeneous groups of non-forensic research volunteers. The inclusion of only M^- cases puts the credibility of the average correlation of $r = 0.31$ as a valid estimation of information overlap, as well as the independence of validity tests as predictors of malingering status in a *U* consisting of both M^+ and M^- cases, into serious question.

Following the publication of the meta-analysis (Berthelson et al., 2013), another study has examined information overlap among a set of 12 validity tests in a mixed malingering-status sample (Meyers et al., 2014). The reported finding was that these 12 validity tests are independent predictors of malingering status. In this study, Pearson product-moment correlations among PVTs reportedly "… ranged from -. 041 … to 0.478…," and the "average of the correlations" was reported as $r = 0.123$ (p. 228). These correlation coefficients are characterized by the authors as "… not found to be statistically significantly different from zero," despite the reported $r = 0.478$ at the stated sample size of $N = 255$ being statistically significant at $p < 0.00001$, and $r = 0.123$ at $N = 255$ being significant at $p < 0.05$. A subsequent reanalysis of these data revised the average Pearson product-moment correlation to $r = 0.26$, with a range of $r = -0.077$ to $r = 0.615$ (Larrabee et al., 2019).

But even if research on PVT information overlap employs a mixed malingering status sample, is the Pearson product-moment correlation coefficient a conceptually appropriate statistic to examine degree of independence among PVTs in predicting malingering status? This measure of correlation is appropriate when there are two continuous variables that meet several assumptions. Specifically, each variable has to be normally distributed, the relationship among variables has to be linear, and residual values from predicting one variable from knowledge of the other variable have to be equally distributed across the measurement range (Baak et al., 2020, p. 3; Tabachnick et al., 2019, pp. 67–76). However, PVTs invariably violate these assumptions. The normality assumption is violated because PVTs are typically severely skewed due to a ceiling effect, whereby participants who fall on the PVT^- side of the range often have near identical scores at one extreme of the scale. A typical example comes from one TOMM validation study, which found about 89% of M^- examinees obtained the maximum score of 50 on Trial 2 (Erdodi & Rai, 2017). Statistical evidence for skewness among PVTs in mixed malingering status samples has also been found, with one study reporting an average absolute skewness of 1.00 among 16 PVTs (Larrabee et al., 2019).[9]

[7] This quantity may also be referred to as the conditional correlation of PVT_1 and PVT_2, conditioned on M^+.

[8] This quantity may also be referred to as the unconditional correlation of PVT_1 and PVT_2, that is, the correlation is not conditioned on malingering status.

[9] In the original work (Larrabee et al., 2019, Table 4, page 1362) average absolute skew among 11 validity tests is reported as – .942. This value is arithmetically impossible because the average of absolute values cannot be a negative quantity. Also, while said Table 4 reports skew for 16 validity tests, said average skew computation is based on a subset of only 11 validity tests with no rationale given for exclusion of the other 5 validity tests. Therefore, in the present analysis, the average skew from all 16 validity tests was recalculated based on all 16 skew values as reported in the original work. This is the value given in the text above (average absolute skew = 1.00).

Given the sample size of $N = 255$ in this study, the critical value for skewness that corresponds to $z = 1.96$ is an absolute skewness of > 0.294 at $\alpha < 0.05$ (see Tabachnick et al., 2019, p. 69 for computational method). Using this critical value, 15 of the 16 PVTs examined in the study exhibited significant skewness (Larrabee et al., 2019). For example, skewness for the Immediate Recognition trial in Word Memory Test was $z = 6.27$, and was $z = 4.98$ for Meyer's Index. With skewness of this magnitude, assumptions of linearity in the relationship among validity tests and of equal distribution of residuals are also untenable. This means that even if the association between pairs of PVTs is examined in a mixed-malingering status sample, the Pearson product-moment correlation coefficient would not be an appropriate measure of correlation.

A further conceptual consideration is that validity tests are always treated as binomial predictors in the prediction of malingering status, with outcomes dichotomized using a cutoff score. Therefore, the ability of validity tests to contribute independent information to the prediction of malingering status should be examined after dichotomization using statistics appropriate for binominal data.

The χ^2 statistic is the most widely accepted inferential test of independence between two categorical variables (cf. Baak et al., 2020, p. 4) and is particularly appropriate for examining the independence of diagnostic tests (Collins & Huynh, 2014, p. 10). A significant χ^2 value provides evidence to reject the null hypothesis that two nominal or two binominal dichotomous variables are independent.

If the null hypothesis in a χ^2 test is rejected and significant information overlap among the variables is found, it is desirable to quantify the degree of association between the variables. Unlike the Pearson product-moment correlation, which quantifies the degree of association among two continuous variables, there is no single gold-standard coefficient for the degree of association among two binominal variables. The two most popular statistics available for this purpose are the Pearson tetrachoric correlation (r_{tet}) and the phi-coefficient (cf. Ekström, 2011). Note that the phi-coefficient is a special case of Pearson product-moment correlation when r is computed for two binominal variables (Tabachnick et al., 2019, p. 776). Both the phi-coefficient and r_{tet} value range from – 1 to + 1, with a value of 0 indicating statistical independence and – 1 or + 1 indicating perfect negative or positive correlation.

However, the phi-coefficient has three distinct disadvantages when quantifying the association between two PVTs. First, it only ranges from – 1 to + 1 when cases are equally distributed among the categories. In other words, the possible maximum and minimum values for the phi-coefficient depend on the marginal probabilities of the 2 × 2 cross-classification table for the two PVTs. Theoretically, the maximum possible value in the range may be restricted to as low as 0.25 when there are many true positives, as would be expected when two PVTs both indicate malingering (cf. Ekström, 2011, p. 9). Second, unlike the product-moment correlation coefficient, the value of phi is not linearly related to degree of association when cells are not equally populated (Davenport & El-Sanhurry, 1991). Finally, because of the two previous problems, interpreting the meaning of a phi-coefficient may be nonintuitive for neuropsychological examiners with extensive background interpreting the Pearson product moment-correlation coefficient.

In contrast, r_{tet} is the measure of binominal association that behaves analogously to the Pearson product-moment correlation (cf. Baak et al., 2020; Kaltenhauser & Lee, 1976). Therefore, r_{tet} is the best statistic to quantify degree of association if a significant χ^2 is found between two PVTs. r_{tet} is a good estimator of correlation even when the underlying variables are skewed, especially as sample sizes become larger (cf. Kaltenhauser & Lee, 1976, p. 310). An additional advantage is that r_{tet} can be readily estimated from the OR (Becker & Clogg, 1988; Bonett, 2007; Digby, 1983) which, in turn, can be calculated not only from hypothetical simulation data (see the next section), but also from the data reported in prior validation studies, such as those reanalyzed below (see Table 2). Because marginal *n*s are typically uneven in malingering studies, and hence in many simulation data matrices that model PVTs, the variant of the r_{tet} formula from Digby (1983) was used in this paper (see Table 2) because it gives a better estimate of r_{tet} than Pearson's original formula when marginal *n*s are unequal (Bonett, 2007).

Statistical Simulation of Mathematically Possible Statistical Independence Among PVTs

To date, there do not appear to be studies where the independence of validity tests has been evaluated in the manner proposed above. Therefore, use of these statistics was first explored in a statistical simulation using a hypothetical data set simulating 100 hypothetical forensic examinees with a base rate of malingering of 0.4. The simulation also includes two PVTs. Each predicts malingering status with sensitivity of 0.5 and specificity of 0.9. These PVT operating characteristics are commonly observed for widely used PVTs in the prediction of malingering (cf. Sherman et al., 2020, p. 14). These parameters yield 60 hypothetical examinees who are M^- and 40 who are M^+. Among the 60 M^- examinees, each PVT gives six false positive and 54 true negative indications. Among the 40 M^+ examinees, each PVT gives 20 true positive and 20 false negative indications. To compute the degree of independence between these PVTs as predictors of malingering, the question arises as to how many indications are concordant between PVT_1 and PVT_2 and how many are discordant. At the upper extreme, all 100 pairs of indications could be concordant between PVT_1 and PVT_2. At the lower extreme, as few as 48 pairs of indications could be

Table 2 Evaluation of information overlap from Test of Memory Malingering (TOMM) and Word Memory Test (WMT) studies reporting association between two PVTs in a mixed malingering status sample

	PVT_1	$n(PVT_1-/PVT_2+)$	$n(PVT_1+/PVT_2\text{-})$	$n(PVT_1-/PVT_2+)$	$n(PVT_1-/PVT_2\text{-})$	χ^{2a}	N	Base rate of PVT_2+	Sensitivity of PVT_1 in detecting PVT_2	Specificity of PVT_1 in detecting PVT_2	OR	r_{tet} (Digby, 1983)
Algebraic relationships/ formula		a	b	c	d	Preacher (2001)	a+b+c+d	(a+c)/ (a+b+c+d)	a/(a+c)	d/(b+d)	(a/b)/(c/d)	$(\sqrt[3]{OR}-1)/(\sqrt[3]{OR}+1)$
TOMM studies												
Armistead-Jehle and Gervais (2011), WMT data	WMT	**24**	**2**	**90**	**229**	**41.79*****	345	.33	.21	.99	**30.53**	**.86**
Armistead-Jehle and Gervais (2011), NV-MSVT data	NV-MSVT	24	6	50	265	**63.10*****	345	**.21**	**.32**	**.98**	**21.20**	**.82**
Armistead-Jehle and Gervais (2011), MSVT data	MSVT	**24**	**3**	**45**	**273**	**82.28*****	345	.20	.35	.99	**48.53**	**.90**
Green (2011)	NV-MSVT	15	1	25	203	**68.84*****	244	**.16**	**.38**	**1.00**	**121.80**	**.95**
Davis et al. (2012)	MSVT	**19**	**3**	**12**	**60**	**33.95*****	92	.34	.61	.95	**31.67**	**.86**
Erdodi and Rai (2017), TBI WMT data	WMT	**15**	**0**	**22**	**46**	**20.11*****	84	.45	.41	1.00	**64.07**[b]	**.92**
Erdodi and Rai (2017), TBI NV-MSVT data	NV-MSVT	**10**	**0**	**8**	**66**	**36.49*****	84	.21	.56	1.00	**164.29**[b]	**.96**

Table 2 (continued)

	PVT_1	n(PVT_1 −/ PVT_2+)	n(PVT_1 +/ PVT_2-)	n(PVT_1 −/ PVT_2+)	n(PVT_1 −/ PVT_2-)	χ^{2a}	N	Base rate of PVT_2+	Sensitivity of PVT_1 in detecting PVT_2	Specificity of PVT_1 in detecting PVT_2	OR	r_{tet} (Digby, 1983)
Algebraic relationships/ formula		a	b	c	d	Preacher (2001)	a + b + c + d	(a + c)/ (a + b + c + d)	a/(a + c)	d/(b + d)	(a/b)/(c/d)	$(\sqrt[3]{OR} - 1)$ $/(\sqrt[3]{OR} + 1)$
Erdodi and Rai (2017), TBI EI-5 data	EI-5	**10**	**0**	**20**	**54**	**17.38*****	84	.36	.33	1.00	**55.83**[b]	**.91**
Erdodi and Rai (2017), psychiatric WMT data	WMT	**6**	**0**	**12**	**50**	**14.37*****	68	.27	.33	1.00	**52.52**[b]	**.90**
Erdodi and Rai (2017), psychiatric NV-MSVT data	NV-MSVT	**4**	**2**	**9**	**53**	**6.55***	68	.19	.31	.96	**11.78**	**.73**
Erdodi and Rai (2017), psychiatric EI-5 data	EI-5	**4**	**2**	**9**	**53**	**6.55***	68	.19	.31	.96	**11.78**	**.73**
Greiffenstein et al. (2008), asymmetrical TOMM scoring data	WMT	101	3	129	240	**123.00*****	473	**.49**	**.44**	**.99**	**62.64**	**.91**
Oudman et al. (2020), VAT-E DR	VAT-E DR	2	0	0	18	**10.43****	20	**.10**	**1.00**	**1.00**	**185**[b]	**.96**
Oudman et al. (2020), VAT-E CI	VAT-E CI	2	0	0	18	**10.43****	20	**.10**	**1.00**	**1.00**	**185**[b]	**.96**

Table 2 (continued)

	PVT_1	$n(PVT_1-/$ $PVT_2+)$	$n(PVT_1+/$ $PVT_2-)$	$n(PVT_1-/$ $PVT_2+)$	$n(PVT_1-/$ $PVT_2-)$	χ^{2a}	*N*	Base rate of PVT_2+	Sensitivity of PVT_1 in detecting PVT_2	Specificity of PVT_1 in detecting PVT_2	OR	r_{tet} (Digby, 1983)
Algebraic relationships/ formula		a	b	c	d	Preacher (2001)	a + b + c + d	(a + c)/ (a + b + c + d)	a/(a + c)	d/(b + d)	(a/b)/(c/d)	$(\sqrt[3]{OR}-1)$ $/(\sqrt[3]{OR}+1)$
WMT study												
Meyers et al. (2014)[c]	FC	**71**	**1**	**27**	**156**	**150.04*****	255	.38	.72	.99	**410.22**	**.98**
	DL	**65**	**5**	**33**	**152**	**117.64*****			.66	.97	**59.88**	**.91**
	SR	**62**	**1**	**36**	**156**	**123.88*****			.63	.99	**268.67**	**.97**
	JL	**63**	**1**	**35**	**156**	**126.66*****			.64	.99	**280.80**	**.97**
	TT	**61**	**1**	**37**	**156**	**121.12*****			.62	.99	**257.19**	**.97**
	AV	**61**	**4**	**37**	**153**	**110.10*****			.62	.97	**63.06**	**.91**
	FTD	**72**	**6**	**26**	**151**	**134.59*****			.73	.96	**69.69**	**.92**
	MEP	**66**	**1**	**32**	**156**	**135.19*****			.67	.99	**321.75**	**.97**
	RDS	**65**	**8**	**33**	**149**	**107.74*****			.66	.95	**36.69**	**.87**
	WMTMNB	**80**	**6**	**18**	**151**	**159.98*****			.82	.96	**111.85**	**.94**
	MI	**98**	**58**	**0**	**99**	**98.38*****			1.00	.63	**335.07**[b]	**.97**

Normal font: quantity as reported in publication; bold font: quantity calculated based on algebraic relationships or statistical formula

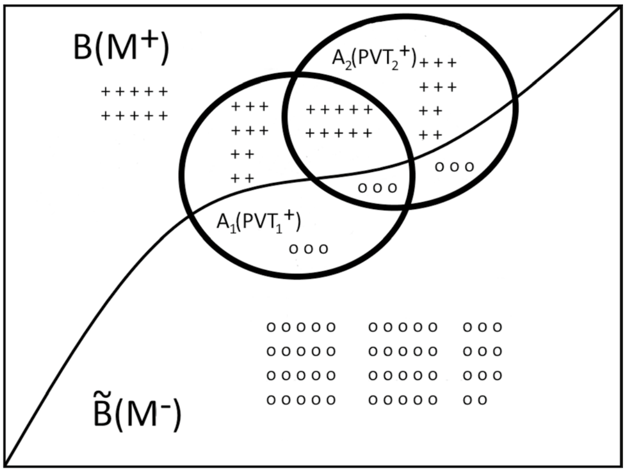
TOMM Test of Memory Malingering (where studies reported classification statistics at various cutoffs for the TOMM data for the M^+/M^- cutoff nearest < 45 was used), *WMT* Word Memory Test, *(NV-)MSVT* (Non-Verbal) Medical Symptom Validity Test, *EI-5* Erdodi-5 Index, *VAT-E DR/CI* Visual Association Test-Extended Delayed Recognition/Consistency Index, *FC* Forced Choice, *DL* Dichotic Listening, *SR* Sentence Repetition, *JL* Judgment of Line Orientation, *TT* Token Test, *AV* Rey Auditory Verbal Learning Test – Recognition, *FTD* Finger Tapping Dominant Hand, *MEP* Memory Error Pattern, *RDS* Reliable Digit Span, *WMTMNB* Word Memory Test – Meyers Neurological Battery, *MI* Meyers Index from MMPI-2 or MMPI-2-RF, *PVT* performance validity test, r_{tet} tetrachoric correlation, across all 23 datasets: weighted by Ns, average r_{tet} .92

*$p < .05$; **$p < .01$; ***$p < .001$

[a] Yates corrected

[b] Quantities a, b, c, and d increased by .5 for computation of OR (cf. Glas et al., 2003)

[c] Larrabee et al. (2019) present a reanalysis of data from Meyers et al. (2014) which includes the claim that in the original work, sensitivity was mistaken for positive predictive power (PPP) and specificity for negative predictive power (NPP) (Larrabee et al., 2019, pp. 1369, first paragraph). Because sensitivity and PPP and specificity and NPP are interchangeable in the formula for the OR on which computation of r_{tet} is based (cf. Ostrowski & Ostrowski, 2020, p. 2), this claim has no bearing on the computations in this table, whether or not Larrabee et al. (2019) are correct



Note. M^{+}= positive for malingering; M^{-}= negative for malingering; $B\ (M^{+})$ = Cases that are M^{+}; $\widetilde{B}\ (M^{-})$ = Cases that are M^{-}; + = a single case that is M^{+}; o = a single case that is M^{-}; PVT_1^{+}= a positive indication of malingering on the first validity test; PVT_2^{+}= a positive indication of malingering on the second validity test; $A_1(PVT_1^{+})$ = cases that are PVT_1^{+}; $A_2(PVT_2^{+})$ = cases that are PVT_2^{+}. Not explicitly shown is that all cases not in the circle of $A_1(PVT_1^{+})$ are $\tilde{A}_1(PVT_1^{-})$ and all cases not in the circle of $A_2(PVT_2^{+})$ are $\tilde{A}_2(PVT_2^{-})$.

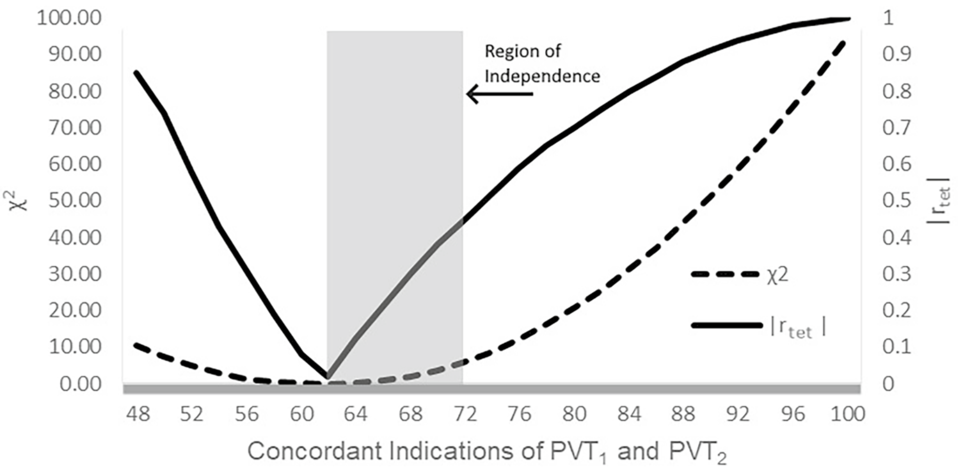
Fig. 4 Hypothetical universe of 100 forensic cases that are either malingering or not malingering ($M^{+} \cap M^{-} = \varnothing$). Base rate of M^{+}= .4. Showing prediction of malingering status based on knowledge of two performance validity tests (PVT_1 and PVT_2), each PVT with sensitivity = 0.5 and specificity = 0.9. PVT_1 and PVT_2 are not independent ($\chi.^2$=8.90, p < 0.01) and are correlated at r_{tet}=0.52. The 74 concordant indications are proportionately equally concordant for M^{+} and M^{-} cases

concordant. This would be the case if both PVTs were perfectly discordant on M^{+} indications and all six false positive indications among M^{-} examinees were simultaneously also discordant, which would limit concordances to only the 48 remaining concordant true negative PVT^{-} indications among the remaining M^{-} cases. In between these upper and lower extremes are 25 possible intermediate concordance scenarios with 98, 96, 94, …., 54, 52, and 50 concordant M^{+} indications for a total of 27 possible concordance scenarios for pairs of PVTs with the typical operating characteristics. For visualization, Fig. 4 depicts an intermediate scenario with 74 concordant indications with proportionately equally concordant indications for M^{+} and M^{-}.

Simulation data were created for all 27 concordance scenarios and degree of statistical independence between PVT_1 and PVT_2 was evaluated using χ^2. A one-tailed test of significance was chosen such that only concordance scenarios with positive correlations among PVT_1 and PVT_2 could be found significant. This choice was made because negative PVT correlations would indicate that a finding of PVT_1^{+} was a significant predictor of a finding of PVT_2^{-}. This outcome, however, would contradict the requirements of malingering determination guidelines that call for a determination of M^{+} to be based on two or more PVTs scoring PVT^{+} in the same direction (cf. Larrabee et al., 2007; Sherman et al., 2020; Slick et al., 1999). Results are presented in Fig. 5.

This statistical simulation shows that, under the baseline, specificity, and sensitivity conditions typically seen in the PVT validation literature, two PVTs may be statistically independent if there are between 62 and 72% concordances. Therefore, hypothetically, it is statistically possible for two PVTs that possess the operating characteristics typical of PVTs to be statistically independent when used in a neuropsychological evaluation. However, this is true for only a narrow region of concordances. The important question that remains is whether data from prior PVT validation studies support the premise that actual PVT associations fall into that region.

Fig. 5 χ2 and absolute values of tetrachoric correlation (r_{tet}) for two hypothetical validity tests (PVT_1 and PVT_2, each with sensitivity of 0.5 and specificity of 0.9, in predicting malingering status (base rate = 0.4), among hypothetical forensic examinees (N = 100) calculated from hypothetical simulation data for all possible permutations of concordant malingering indications



Note: Region of Independence shows where null hypothesis of independence of PVT_1 *and* PVT_2 can NOT be rejected (p >.05, one tailed; 62 concordances to 72 concordances). The r_{tet} values for concordances ranging from 48 to 60 concordances are negative, those for 62 to 100 concordances are positive.

Reanalysis of Data from Prior PVT Validation Studies

This question was examined through a reanalysis of data from validation studies reporting associations between two validity tests in a mixed malingering-status sample. The reanalysis focused on associations involving the TOMM and the WMT. The TOMM and WMT were chosen because these are the two most commonly used PVTs in neuropsychological testing (Martin et al., 2015, p. 762). Independence of, and correlations between TOMM-based malingering classification and malingering classification based on four other validity tests were reanalyzed using data from all TOMM validation studies reported in a recent systematic review and meta-analysis (Martin et al., 2020). In these studies, the TOMM was used in a mixed malingering status-sample together with at least one other validity test and a 2 × 2 cross classification table of the TOMM and the second validity test was either provided, or enough other information was reported to allow algebraic imputation of such a table. Studies were selected by Martin et al. (2020) following a well-defined and well-documented selection process (pp. 90–94) and exhaustively cover the period from 1997, the year the TOMM was first published, to approximately mid-2019. No attempt was made to search for additional TOMM studies published after that cutoff date.

Additionally, to cross-check the reanalysis of TOMM data, data from one malingering classification study using the WMT with a mixed malingering-status sample was similarly reanalyzed (Meyers et al., 2014; Table 3). This was the same WMT study that provided data for the reanalysis reported by Larrabee et al. (2019). Results are reported in Table 2.

Examination of the 25 paired associations among these 18 validity tests (see Table 2) shows that all the PVTs provide highly overlapping information. All χ^2 statistics were significant, showing a lack of independence between PVTs, and the average bivariate tetrachoric correlation, weighted by study sample size, was $r_{\mathrm{tet}} = 0.92$, which is dramatically higher than the correlations reported in prior work (Berthelson et al., 2013). All bivariate correlations were well above the cutoff of 0.7 for including two variables as predictors of the same outcome, and most correlations were well above the cutoff of 0.9, where variables are considered fully collinear and thus redundant predictors (cf. Tabachnick et al., 2019, p. 77). The degree of association between PVTs is in the order of the short-term test-test reliability of well validated cognitive tests, such as the Wechsler Adult Intelligence Scale (cf. Schuerger & Witt, 1989). While these studies report correlations between pairs, the present correlational reanalysis also implies that all 18 validity tests are redundant. This is because, by virtue of the transitive property of the law of equality, when redundancy is shown among a group of tests, hypothetically, between PVT_1 and PVT_2 and PVT_2 and PVT_3, then PVT_1 and PVT_3 must also be redundant. The correlations found in the present reanalysis of the data from Meyers et al. (2014) differ from those found in the reanalysis by Larrabee et al. (2019) because in the present analysis, independence and correlation were examined subsequent to dichotomization, while Larrabee et al. (2019) calculated the correlations using the raw scores. Furthermore, in the present analysis, the Pearson tetrachoric coefficient of correlation (r_{tet}) is used, while Larrabee et al. used the Pearson product-moment correlation coefficient, despite the assumptions of the test being violated due to skewness (see discussion above).

What might explain collinearity of this magnitude? Legal and ethical test security requirements preclude a detailed presentation of this argument. Suffice to say that many PVTs, such as the TOMM, RDS, MSVT, PDRT, WMT, or Rey 15-Item Test (FIT; Millis & Kler, 1995; Rey, 1964), are close variants of a prototypical test theme where, upon having been serially presented with simple stimuli, examinees are challenged with recognition trials, free recall trials, or both. These tests vary only slightly in the type of simple stimulus used, such as words, letters, numbers, drawings, symbols, or a mixture of these. Other PVTs, such as the Finger Tapping Dominant Hand Test (Meyers & Volbrecht, 2003), involve motoric responses and may be redundant with memory tests due to common factors such as attention or illness perception (cf. Henry et al., 2018).

In sum, commonly administered PVTs likely provide information that is substantially redundant toward determining malingering status, and are thus collinear. Researchers and forensic examiners using more than one PVT in the determination of malingering status would do well to accurately estimate the independence of PVTs before predicting malingering status based on more than one PVT. To avoid including collinear PVTs, information overlap among PVTs should be calculated using methods appropriate for estimating independence and correlation between categorical variables subsequent to the dichotomization of PVT outcomes and including participants of all malingering-status classifications. PVTs found to correlate at $r_{\mathrm{tet}} > 0.7$ should be combined, and if PVTs correlate at $r_{\mathrm{tet}} > 0.9$, redundant PVTs should be eliminated. Estimates of PVT correlations obtained from homogeneous malingering groups where all participants were either exclusively M^+ or exclusively M^- should not be considered when estimating collinearity and information overlap among PVTs as predictors of malingering status.

Evaluation of Classification Accuracy of PVT Aspects of the Malingering Determination Algorithms Compared to Alternatives

Malingering determination algorithms are thus limited by the minimal additional information contributed to the determination of malingering status from PVTs beyond the first. Regardless, all determination algorithms (Larrabee et al., 2007; Sherman et al., 2020; Slick & Sherman, 2012; Slick et al., 1999), as well as Consensus Conference Statements from the American Academy of Clinical Neuropsychology (Heilbronner et al., 2009, p. 12; Sweet et al., 2021, p. 1067), call for the simultaneous consideration of multiple PVTs in the determination of malingering status. All algorithms agree a logical "and" (probabilistically multiplicative) combination of PVTs is needed when scores from multiple PVTs are considered. Any PVT^- results are to be disregarded, with only PVT^+ results being considered, as long as the ratio of PVT^+ to PVT^- findings is at least 2 to 5 (Chafetz, 2020; Sherman et al., 2020; Sweet et al., 2021).

Various rationales are offered for this method of combining information from multiple PVTs, including "Positive Likelihood Chaining," a pseudo-Bayesian method which superficially appears similar to the simple Bayes method, also known as "Naïve Bayes," "Independence Bayes," or "Idiot's Bayes," which is itself recognized both as mathematically incorrect and unrelated to Bayes theorem unless predictors are conditionally independent (Hand & Yu, 2001, p. 386; Steyerberg, 2009, p. 63; also see detailed discussion of use and misuse of Bayes' Theorem in the determination of malingering in Supplemental Appendix A). Briefly, an actual simple Bayes method would work as follows. After ascertaining that all included PVTs are independent conditional on malingering status, using all PVT scores, depending on whether a PVT was PVT^+ or PVT^-, each corresponding coefficient of the likelihood ratio for positive results or likelihood ratio for negative results (the likelihood ratio for positive results if the PVT is PVT^+, the likelihood ratio for negative results if it is PVT^-) would be multiplied with the odds of the base rate of malingering to compute the posterior odds of malingering. This method is mathematically incorrect unless predictors are conditionally independent and is tantamount to creating a logistic regression equation with multiple predictors where all beta weights are obtained from single variable prediction (Steyerberg, 2009, p. 64; Zadora et al., 2014, p.92). With predictors that lack conditional independence, this simple Bayes method repeatedly counts any overlapping variance among predictors as many times as there are predictors in the equation. The simple Bayes method has therefore been shown to overestimate predictability (Hand & Yu, 2001, p. 388; Zadora et al., 2014, p. 209). As shown in Supplemental Appendix A, PVTs with the typical operating characteristics (sensitivity = 0.5 and specificity = 0.9) cannot be independent conditioned on malingering status. The simple Bayes method is therefore not tenable to combine findings from multiple PVTs in the determination of malingering status, but there does not appear to be a prior published attempt to do so.

However, positive likelihood chaining has been proposed as a suitable method to combine findings from multiple PVTs in the determination of malingering (cf. Chafetz, 2011, 2020; Larrabee, 2008; Lippa, 2018; Meyers et al., 2014). This method goes one step beyond simple Bayes and disregards any PVT^- findings from the simple Bayes equation, with the posterior odds calculated as the product of the pretest odds of malingering and the likelihood ratios for positive results from all PVTs with a finding of PVT^+ (cf. Chafetz, 2011, 2020; Larrabee, 2008; Lippa, 2018; Meyers et al., 2014). While the pseudo-Bayesian rationale that seemingly justifies positive likelihood chaining is not universally accepted in the malingering literature (see Larrabee et al., 2019 for a discussion), it does provide a pseudo-mathematical rationale for disregarding PVT^- outcomes when evaluating malingering criteria related to PVTs, as long as the ratio of PVT^+ to PVT^- findings is at least two to seven as stipulated in the malingering determination algorithm (cf. Sherman et al., 2020). Please see Supplemental Appendix A for a detailed analysis of the mathematical untenability of positive likelihood chaining as a method to combine findings from multiple PVTs in the determination of malingering status.

How, then, does classification accuracy of the PVT components of the malingering determination algorithm (two or more of seven PVTs must be PVT^+ to determine M^+) compare to that of other possible prediction algorithms, such as a base rate (constant only) model, prediction from one PVT, or prediction based on logistic regression? This question will be addressed through an experimental evaluation of these classification algorithms, quantifying their ability to correctly classify malingering status (cf. Webb, 2017). Experimental evaluation of diagnostic algorithms is often performed with actual data sets where both the diagnostic indicators and the condition of interest are known (e.g., Macek-Jilkova et al., 2021; Silva & Bernardino, 2022). However, when an actual data set is not available, as is the case here, researchers in neuropsychology use hypothetical data sets to evaluate the performance of classification algorithms (e.g., Gates et al., 2016; Underwood et al., 2018), which is the approach taken here.

To quantify the classification accuracy of malingering determination algorithms, five hypothetical data sets were thus created, each with 100 forensic examinees with a base rate of M^+ of 0.4. These data sets also included seven hypothetical PVTs (PVT_1 through PVT_7), each with a sensitivity of 0.5 and a specificity of 0.9 for determining malingering status. Performance of predictive algorithms cannot be compared for sets of seven PVTs with these sensitivities and specificities that are also simultaneously unconditionally or conditionally independent predictors of malingering status, because it is not mathematically possible for a set of more than two PVTs to exhibit all three attributes simultaneously (a sensitivity of 0.5, specificity of 0.9, and that are statistically unconditionally independent of each other), and no two PVTs can be conditionally independent conditioned on M status (see Supplemental Appendix A for detailed explanation). Note also that the PVT algorithm yields unacceptably low operating characteristics when malingering status is determined from two statistically unconditionally independent PVTs (combined sensitivity = 0.08, specificity = 0.93, diagnostic OR = 1.15, and posterior probability or positive predictive power in a forensic setting = 0.43; see Supplemental Appendix A for detailed calculations). For these reasons, a comparison of predictive algorithms will be modelled for five hypothetical simulation data sets that reflect relationships among seven PVTs that *are* mathematically possible:

1. In the first three hypothetical simulation data sets, the mean Pearson r value for the 36 paired correlations among the seven *PVTs* is $r \approx 0.31$ (cf. Berthelson et al., 2013). This is the average correlation among PVTs reported in the malingering literature, albeit possibly erroneously (see above). Further, modelled on the TOMM, each hypothetical PVT has scores ranging from 39 to 50 with the cutoff for M^+ at PVT < 45 (range 39–44), with higher scores indicating M^- and ranging from 49 to 50. Additionally, the question arises as to how to handle concordant M^+ indications that are key to the malingering algorithm, or specifically, how PVT^+ concordances should be distributed among M^+ and M^- cases. A correlation of $r \approx 0.31$ allows for approximately 40 to 46 concordant PVT^+ indications among the 100 simulation cases. The present simulation will model three possible concordance scenarios, including (a) a best-case scenario, with 40 PVT^+ indications based on two or more concordant PVTs only among M^+ cases, yielding 40 true positive and zero false positive cases, (b) a "middle-of-the-road" scenario, with 37 PVT^+ concordances among M^+ cases and six PVT^+ concordances among M^- cases, yielding 37 true positive cases and six false positive cases, and (c) a worst-case scenario, with 34 PVT^+ concordances among M^+ cases and 12 PVT^+ concordances among M^- cases, yielding 34 true positive cases and 12 false positive cases.
2. In the fourth and fifth hypothetical simulation data sets, the average correlation among all PVT pairs is $r_{\text{tet}} \approx 0.92$ after dichotomization, which is the average correlation among the 18 validity tests from the reanalysis in Table 2. Here as well, considered individually, all seven PVTs are predictors of malingering, each with sensitivity of 0.5 and specificity of 0.9. An average correlation of $r_{\text{tet}} \approx 0.92$ among the PVT pairs allows for approximately 30 to 32 concordant M^+ indications among the 100 simulation cases. Here, two concordance scenarios are modelled, including (a) a best-case scenario, with 26 PVT^+ indications based on two or more concordant PVTs among M^+ cases and six PVT^+ indications among M^- cases, yielding 26 true positive and six false positive cases, and (b) a worst-case scenario with 20 PVT^+ indications based on two or more concordant PVTs among M^+ cases, and 12 PVT^+ indications among M^- cases, yielding 20 true positive and 12 false positive cases.

The classification accuracy of the PVT aspect of malingering determinations algorithms (Larrabee et al., 2007; Sherman et al., 2020; Slick & Sherman, 2012; Slick et al., 1999) will be compared under both correlation conditions with alternative classification algorithms, including (a) a constant only model, (b) prediction based on a single PVT only, and (c) prediction based on direct logistic regression using all seven PVTs as predictors of malingering status. Results of the simulation are presented in Table 3. Note that 95% confidence intervals were not computed as this was only a simulation using hypothetical data to compare algorithm performance, not an estimation of population parameters from sample data.

When PVTs are correlated at $r \approx 0.31$, the simulation shows that the classification accuracy of PVT aspects of the malingering determination algorithm (Larrabee et al., 2007; Sherman et al., 2020; Slick et al., 1999; Slick & Sherman, 2012), and of logistic regression, depends on how the concordances fall. In the best-case scenario, with concordant PVT^+ scores exclusively among M^+ cases, both algorithms perform best, although the malingering algorithm outperforms logistic regression for both sensitivity and specificity. When concordant PVT^+ scores also occur for some M^- cases, specificities of both algorithms remain ≥ 0.9, but the malingering algorithm has better sensitivity. However, in the worst-case scenario where PVT^+ concordances occur more frequently among M^- cases, while the malingering algorithm still has better sensitivity, specificity drops to 0.8, which is below the minimum of 0.9 required in the malingering determination algorithm, even before considering 95% confidence intervals.

When two PVTs are correlated at $r_{\text{tet}} \approx 0.92$, due to PVT redundancy, logistic regression reverts to single variable prediction because redundant PVT predictors are unable to improve prediction over what is possible based on knowledge from a single PVT. In the best-case concordance scenario, PVT-specific rules from the malingering determination algorithm (Larrabee et al., 2007; Sherman et al., 2020; Slick et al., 1999; Slick & Sherman, 2012) maintain specificity at 0.9, but improve sensitivity to 0.6 over the single PVT sensitivity of 0.5. However, in the worst-case scenario, specificity for the malingering algorithm again drops below the required minimum of 0.9 to 0.8, while sensitivity is unimproved over single PVT prediction at 0.5.

In sum, in both correlation scenarios, the malingering determination algorithm for combining findings from multiple PVTs (two of seven PVTs need to be PVT^+ to

Table 3 Comparison of classification accuracy of various malingering (M) prediction algorithms in a universe of 100 hypothetical forensic examinees with base rate of M + of .4

Predictors	True +	False +	False −	True −	Sensitivity	Specificity	% Correct	True +	False +	False −	True −	Sensitivity	Specificity	% Correct
	M prediction based on constant only													
Constant only	0	0	40	60	0	1	60							
	M prediction based on cross classification of single PVT with M													
Single PVT	20	6	20	54	.50	.90	74							
7 PVTs mean $r \approx .31$[a]	M prediction with direct logistic regression of all 7 PVTs on M							M prediction: M + if 2 or more of 7 PVTs indicate M +						
CON: 40 M + /0 M-[b]	39	1	1	59	.98	.98	98	40	0	0	60	1	1	100
CON: 37 M + /6 M-[c]	36	6	4	54	.90	.90	90	37	6	3	54	.93	.90	91
CON: 34 M + /12 M-[d]	31	7	9	53	.78	.88	84	34	12	6	48	.85	.80	82
7 PVTs mean $r_{tet} \approx .92$[a]														
CON: 24 M + /6 M-[e]	reverts to single PVT model due to PVT redundancy							24	6	16	54	.60	.90	76
CON: 20 M + /12 M-[f]								20	12	20	48	.50	.80	68

PVT performance validity test, *M* malingering, *CON* concordance of two or more *PVTs* indicating *M* + status, r_{tet} tetrachoric correlation coefficient

[a] PVT_1 through PVT_7 each with sensitivity = .5 and specificity = .9 for predicting M, average r or r_{tet} values were calculated from all 36 bivariate correlations among the 7 PVTs

[b] Actual mean $r = .29$ (range .14–.46)

[c] Actual mean bivariate $r = .32$ (range .14–.50)

[d] Actual mean $r = .34$ (range .15–.54)

[e] Actual mean bivariate $r_{tet} = .92$ (range .88–.96)

[f] Actual mean bivariate $r_{tet} = .91$ (range .88–.96)

determine M^{+}) consistently favors sensitivity over specificity compared to decisions about malingering status based on logistic regression. While logistic regression consistently maintains specificity at or near the target level of 0.9, the malingering algorithm yields specificities well below 0.9 under less-than-ideal concordance scenarios. Because it is unclear which concordance scenarios occur when testing actual forensic examinees, classification based on logistic regression would therefore appear to be more conservative. Had 95% confidence intervals been calculated, the lower bounds of all classification accuracy estimates would have been even lower (see discussion above).

When aggregating findings from multiple PVTs in the determination of malingering status, it is therefore vital that collinearity (or multicollinearity) and information overlap among PVTs be evaluated first, and that redundant PVTs be eliminated. Then, either a single PVT model, or if multiple non-redundant PVTs are available, a model based on logistic regression promises the best compromise between maintaining specificity at or near 0.9 while optimizing classification accuracy[10] (cf. Bossuyt et al., 2013; Victor et al., 2009; Wolfe et al., 2010).

Logistic regression also has the advantage of providing excellent diagnostics for the detection of collinearity (or multicollinearity) among predictor PVTs, in case the initial PVT selection inadvertently included PVTs with significant information overlap (cf. Midi et al., 2010). An additional advantage of logistic regression is that it can address overfitting concerns through the use of bootstrapping methods (see Steyerberg et al., 2001) to examine whether findings from PVT validation studies can be cross-validated (cf. de Rooij & Weeda, 2020).

Based on the above calculations, it may be that the operating characteristics of resulting prediction models will be insufficient to identify malingering with sufficient precision. Because the posterior probability or the posterior odds adds vital information for a forensic neuropsychologist evaluating the likelihood of malingering in a single examinee with a positive determination of malingering, a single likelihood ratio for positive results obtained from the confusion table resulting from the entire regression equation would help to quantify the likelihood of a correct determination (cf. Fischer et al., 2003, p. 1047; Glas et al., 2003, p. 1134; Moons et al., 2012, p. 1411).

Widespread adoption of a logistic regression-based algorithm would require standardizing the administration of PVTs during forensic neuropsychology evaluations. Standardization, however, may not present an insurmountable obstacle, as use of such standardized batteries has already been reported by many forensic clinician-researchers (e.g., Chafetz, 2008, p.532; D. Green et al., 2012, p. 185; Greve et al., 2006a, p. 443; Greve et al., 2006b, p. 1180), with others reporting standardized administration of PVTs as part of a flexible battery approach (e.g., Armistead-Jehle & Gervais, 2011, p. 285; Buddin et al., 2014, p. 529; Rees et al., 1998, p. 16). On the other hand, there has been general resistance to the adoption of standardized batteries in neuropsychology, and their use may facilitate undesirable coaching to "beat" PVTs.

Researchers could also look to biomedical diagnostics where tremendous work has been completed on statistical and methodological issues in the development and validation of multivariable prediction models (e.g., Collins et al., 2015; Moons et al., 2012). This includes, for example, a consensus model on the "transparent reporting of multivariate prediction models for individual prognosis or diagnosis" (TRIPOD; Collins et al., 2015), which includes a checklist of methodological and statistical considerations when developing and evaluating multivariable prediction models, and a review of statistical methods to quantify changes in determinative accuracy when adding putatively predictive tests beyond the first one (Moons et al., 2012). This work is directly applicable to the problem of predicting malingering from knowledge of multiple predictors, including PVTs.

Summary and Brief Discussion of Statistical Issues

In summary, a review of statistical aspects of the literature purporting to validate PVT as predictors of malingering has identified several important issues.

- Studies validating PVTs as predictors of malingering status should always report full confusion table statistics including diagnostic OR and 95% confidence intervals for all statistics.
- When evaluating collinearity among putative predictors of malingering status, such evaluations need to include participants of heterogeneous malingering status. There must be participants that are classified as malingerers, as well as participants that are not classified as malingerers. Studies that evaluate the independence of predictors of malingering only among homogeneous groups of either all malingerers or all non-malingerers should not be considered when evaluating the degree to which pairs of validity scores are collinear.
- It is theoretically possible for two validity tests to be unconditionally independent predictors of malingering status when each PVT has sensitivity of 0.5 and specific-

[10] An additional method for estimating posterior probabilities from multiple PVTs may be Markov Chain Monte Carlo algorithms that correctly implement multivariate Bayesian modelling (cf. Al-Khairullah & Al-Baldawi, 2021). When directly compared to frequentist approaches, such as logistic regression, such models have been shown to yield diagnostic accuracy comparable to logistic regression (e.g., Wang et al., 2014; Witteveen et al., 2018).

ity of 0.9, and the base rate of malingering is 0.4. However, this is true only for a narrow band of concordance scenarios. Once additional PVTs are added beyond the first two, not all PVT pairs can be unconditionally independent, even theoretically. No two PVTs can be conditionally independent conditioned on malingering status.

- Reanalyses of extant data from the malingering literature with χ^2 on paired associations of the TOMM and WMT with a second validity tests show that, of the 18 validity tests examined, none are predictors of malingering status that are statistically independent of other validity tests. In essence, this means that all 18 validity tests are slightly different versions the same test.
- The average r_{tet} of the TOMM and WMT with a second validity test is about 0.92, which means all 18 validity tests that were examined provide redundant information toward the prediction of malingering status, and therefore should not be used together when predicting malingering status from knowledge of validity tests.
- The simple Bayes method and positive likelihood chaining method are mathematically erroneous to combine knowledge from two or more PVTs in the prediction of malingering status.
- When testing the PVT aspect of the malingering determination algorithm (cf. Larrabee, 2008; Larrabee et al., 2007; Sherman et al., 2020; Slick et al., 1999; Slick & Sherman, 2012) (two of seven PVTs have to show PVT^+ to determine an examinee is malingering) against alternative approaches such as prediction from base rate only, single PVT prediction, and prediction using logistic regression, under reasonable association and concordance scenarios, these algorithms, including the "Slick" algorithm, perform unexpectedly depending on where the concordances fall. These algorithms may not yield the goal of maintaining specificity at or above 0.9. Alternatives to the "Slick" algorithm, and related malingering determination algorithms that do not have this particular problem, are single PVT prediction or prediction based on logistic regression. However, these alternative approaches may not be practicable either, as will be further discussed in the companion method review in Part II.
- Future research on malingering prediction models involving multiple validity tests should follow the TRIPOD consensus model (Collins et al., 2015).

Further implications of these statistical issues, as well as implications for forensic practice and future research, are comprehensively discussed in the summary, discussion, and recommendations section and the end of Part II, the methods review, which accompanies this statistics review.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11065-023-09601-7.

Author Contribution Not applicable because this manuscript has only a single author.

Availability of Data and Materials No parts of this work have been presented elsewhere. However, in compliance with data sharing requirements, data and analysis presented in Fig. 4 and Table 3 have been publicly shared under the following Figshare DOIs: Fig. 4 Data and Analysis (chi square and r_{tet} for all 2-validity-measure concordances), https://doi.org/10.6084/m9.figshare.14256470.v2; Table 3 Data and Analysis (Prediction of malingering from 7 validity measures), https://doi.org/10.6084/m9.figshare.14368727.v3.

Declarations

Ethics Approval Not applicable because this is a review and not a study where human or animal data were collected.

Competing Interests The author declares no competing interests.

References

Aita, S. L., Borgogna, N. C., Aita, L. J., Ogden, M. L., & Hill, B. D. (2020). Comparison of clinical psychologist and physician beliefs and practices concerning malingering: Results from a mixed methods study. *Psychological Injury and Law, 13*(3), 246–260. https://doi.org/10.1007/s12207-020-09374-x

Al-Khairullah, N. A., & Al-Baldawi, T. H. K. (2021). Bayesian computational methods of the logistic regression model. *Journal of Physics: Conference Series, 1804*(1), 012073. https://doi.org/10.1088/1742-6596/1804/1/012073

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR.* (4th ed., text revision.). American Psychiatric Association.

Armistead-Jehle, P., & Gervais, R. O. (2011). Sensitivity of the test of memory malingering and the Nonverbal Medical Symptom Validity Test: A replication study. *Applied Neuropsychology, 18*(4), 284–290. https://doi.org/10.1080/09084282.2011.595455

Ashendorf, L., O'Bryant, S., & McCaffrey, R. (2003). Specificity of malingering detection strategies in older adults using the CVLT and WCST. *The Clinical Neuropsychologist, 17*(2), 255–262. https://doi.org/10.1076/clin.17.2.255.16502

Ashendorf, L., Constantinou, M., & McCaffrey, R. J. (2004). The effect of depression and anxiety on the TOMM in community-dwelling older adults. *Archives of Clinical Neuropsychology, 19*(1), 125–130. https://doi.org/10.1016/S0887-6177(02)00218-4

Baak, M., Koopman, R., Snoek, H., & Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics & Data Analysis*, *152*, 107043. https://doi.org/10.1016/j.csda.2020.107043

Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various digit span scores in the detection of suspect effort. *The Clinical Neuropsychologist, 20*(1), 145–159. https://doi.org/10.1080/13854040590947362

Bashem, J. R., Rapport, L. J., Miller, J. B., Hanks, R. A., Axelrod, B. N., & Millis, S. R. (2014). Comparisons of five performance validity indices in bona fide and simulated traumatic brain injury. *The Clinical Neuropsychologist, 28*(5), 851–875. https://doi.org/10.1080/13854046.2014.927927

Bayman, E. O., & Dexter, F. (2021). Multicollinearity in logistic regression models. *Anesthesia and Analgesia, 133*(2), 362–365. https://doi.org/10.1213/ANE.0000000000005593

Becker, M. P., & Clogg, C. C. (1988). A note on approximating correlations from odds ratios. *Sociological Methods and Research, 16*(3), 40–424. https://doi.org/10.1177/0049124188016003003

Benitez-Silva, H., Buchinsky, M., & Rust, J. (2004). *How large are the classification errors in the social security disability award process?* (No. w10219; p. w10219). National Bureau of Economic Research. https://doi.org/10.3386/w10219

Berthelson, L., Mulchan, S. S., Odland, A. P., Miller, L. J., & Mittenberg, W. (2013). False positive diagnosis of malingering due to the use of multiple effort tests. *Brain Injury, 27*(7–8), 909–916. https://doi.org/10.3109/02699052.2013.793400

Bianchini, K. J., Greve, K. W., & Glynn, G. (2005). On the diagnosis of malingered pain-related disability: Lessons from cognitive malingering research. *The Spine Journal, 5*(4), 404–417. https://doi.org/10.1016/j.spinee.2004.11.016

Binder, L. M. (1993). *Portland Digit Recognition Test manual* (second). L. M. Binder.

Binder, L. M., Larrabee, G. J., & Millis, S. R. (2014). Intent to fail: Significance testing of forced choice test results. *The Clinical Neuropsychologist, 28*(8), 1366–1375. https://doi.org/10.1080/13854046.2014.978383

Binder, L. M., & Willis, S. C. (1991). Assessment of motivation after financially compensable minor head trauma. *Psychological Assessment /, 3*(2), 175–181. https://doi.org/10.1037/1040-3590.3.2.175

Bonett, D. G. (2007). Transforming odds ratios into correlations for meta-analytic research. *American Psychologist, 62*(3), 254–255. https://doi.org/10.1037/0003-066X.62.3.254

Bossuyt, P. M. M., Davenport, C., Deeks, J. J., Hyde, C., Leeflang, M. M. G., & Scholten, R. (2013). Chapter 11 Interpreting results and drawing conclusions. In J. J. Deeks, P. M. M. Bossuyt, & C. A. Gatsonis (Eds.), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9.* The Cochrane Collaboration.

Buddin, W. H., Schroeder, R. W., Hargrave, D. D., Von Dran, E. J., Campbell, B., & C. J., Heinrichs, R. J., & Baade, L. E. (2014). An examination of the frequency of invalid forgetting on the Test of Memory Malingering. *The Clinical Neuropsychologist, 28*(3), 525–542. https://doi.org/10.1080/13854046.2014.906658

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology, 20*(4), 419–426. https://doi.org/10.1016/j.acn.2005.02.002

Chafetz, M. D. (2020). Deception is different: Negative validity test findings do not provide "evidence" for "good effort." *The Clinical Neuropsychologist*, 1–37. https://doi.org/10.1080/13854046.2020.1840633

Chafetz, M. D. (2008). Malingering on the social security disability consultative exam: Predictors and base rates. *The Clinical Neuropsychologist, 22*(3), 529–546. https://doi.org/10.1080/13854040701346104

Chafetz, M. D. (2011). Reducing the probability of false positives in malingering detection of social security disability claimants. *The Clinical Neuropsychologist, 25*(7), 1239–1252. https://doi.org/10.1080/13854046.2011.586785

Chafetz, M. D., Bauer, R. M., & Haley, P. S. (2020). The other face of illness-deception: Diagnostic criteria for factitious disorder with proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 34*(3), 454–476. https://doi.org/10.1080/13854046.2019.1663265

Chafetz, M. D., & Underhill, J. (2013). Estimated costs of malingered disability. *Archives of Clinical Neuropsychology, 28*(7), 633–639. https://doi.org/10.1093/arclin/act038

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD statement. *Journal of Clinical Epidemiology, 68*(2), 112–121. https://doi.org/10.1016/j.jclinepi.2014.11.010

Collins, J., & Huynh, M. (2014). Estimation of diagnostic test accuracy without full verification: A review of latent class methods. *Statistics in Medicine, 33*(24), 4141. https://doi.org/10.1002/sim.6218

Davenport, E. C., & El-Sanhurry, N. A. (1991). Phi/Phimax: Review and synthesis. *Educational and Psychological Measurement, 51*(4), 821–828. https://doi.org/10.1177/001316449105100403

Davis, J. J., Wall, J. R., & Whitney, K. A. (2012). Derivation and clinical validation of consistency indices on the test of memory malingering. *Archives of Clinical Neuropsychology, 27*(7), 706–715. https://doi.org/10.1093/arclin/acs078

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science, 3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. (2000). *The California Verbal Learning Test-Second Edition.* The Psychological Corporation.

Digby, P. G. N. (1983). Approximating the tetrachoric correlation coefficient. *Biometrics, 39*(3), 753–757. https://doi.org/10.2307/2531104

Edens, J. F., Truong, T. N., & Otto, R. K. (2020). Classification accuracy of the rare symptoms and symptom combinations scales of the Structured Inventory of Malingered Symptomatology in three archival samples. *Law and Human Behavior, 44*(2), 167–177. https://doi.org/10.1037/lhb0000361

Ekström, J. (2011). The Phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule debate. *UCLA Department of Statistics Papers*. https://escholarship.org/uc/item/7qp4604r

Erdodi, L. A., & Rai, J. K. (2017). A single error is one too many: Examining alternative cutoffs on Trial 2 of the TOMM. *Brain Injury, 31*(10), 1362–1368. https://doi.org/10.1080/02699052.2017.1332386

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. *Intensive Care Medicine, 29*(7), 1043–1051. https://doi.org/10.1007/s00134-003-1761-8

Fletcher, R. H., Fletcher, S. W., & Fletcher, G. S. (2014). *Clinical epidemiology: The essentials: Vol. 5th edition*. LWW.

Gates, K. M., Henry, T., Steinley, D., & Fair, D. A. (2016). A Monte Carlo evaluation of weighted community detection algorithms. *Frontiers in Neuroinformatics, 10*, 45. https://doi.org/10.3389/fninf.2016.00045

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology, 56*(11), 1129–1135. https://doi.org/10.1016/S0895-4356(03)00177-X

Green, D., Rosenfeld, B., Belfi, B., Rohlehr, L., & Pierson, A. (2012). Use of measures of cognitive effort and feigned psychiatric symptoms with pretrial forensic psychiatric patients. *International Journal of Forensic Mental Health, 11*(3), 181–190. https://doi.org/10.1080/14999013.2012.723665

Green, P. W. (2003). *Green's word memory test for windows: User's manual.* Green's Publishing.

Green, P. W. (2004). *Green's medical symptom validity test (MSVT) for microsoft windows: User's manual.* Green's Publishing.

Green, P. W. (2011). Comparison between the Test of Memory Malingering (TOMM) and the Nonverbal Medical Symptom Validity Test (NV-MSVT) in adults with disability claims. *Applied Neuropsychology, 18*(1), 18–26. https://doi.org/10.1080/09084282.2010.523365

Green, P. W., Allen, L. M., & Astner, K. (1996). *Manual for the Word Memory Test*. Cognisyst N.C.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6*(3), 218–224. https://doi.org/10.1037//1040-3590.6.3.218

Greiffenstein, M. F., Greve, K. W., Bianchini, K. J., & Baker, W. J. (2008). Test of Memory Malingering and Word Memory Test: A new comparison of failure concordance rates. *Archives of Clinical Neuropsychology, 23*(7), 801–807. https://doi.org/10.1016/j.acn.2008.07.005

Greve, K. W., Bianchini, K. J., Black, F. W., Heinly, M. T., Love, J. M., Swift, D. A., & Ciota, M. (2006a). Classification accuracy of the Test of Memory Malingering in persons reporting exposure to environmental and industrial toxins: Results of a known-groups analysis. *Archives of Clinical Neuropsychology, 21*(5), 439–448. https://doi.org/10.1016/j.acn.2006.06.004

Greve, K. W., Bianchini, K. J., & Doane, B. M. (2006b). Classification accuracy of the test of memory malingering in traumatic brain injury: Results of a known-groups analysis. *Journal of Clinical and Experimental Neuropsychology, 28*(7), 1176–1190. https://doi.org/10.1080/13803390500263550

Grimes, D. A., & Schulz, K. F. (2005). Epidemiology 3: Refining clinical diagnosis with likelihood ratios. *The Lancet, 365*(9469), 1500–1505. https://doi.org/10.1016/S0140-6736(05)66422-7

Hand, D. J., & Yu, K. (2001). Idiot's Bayes: Not so stupid after all? *International Statistical Review / Revue Internationale De Statistique, 69*(3), 385–398. https://doi.org/10.2307/1403452

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Participants1, C. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, *23*(7), 1093–1129. https://doi.org/10.1080/13854040903155063

Henry, G. K., Heilbronner, R. L., Suhr, G., & J., Wagner, E., & Drane, D. L. (2018). Illness perceptions predict cognitive performance validity. *Journal of the International Neuropsychological Society : JINS, 24*(7), 735–745. https://doi.org/10.1017/S1355617718000218

Jones, A. (2013). Test of memory malingering: Cutoff scores for psychometrically defined malingering groups in a military sample. *The Clinical Neuropsychologist, 27*(6), 1043–1059. https://doi.org/10.1080/13854046.2013.804949

Kaltenhauser, J., & Lee, Y. (1976). Correlation coefficients for binary data in factor analysis. *Geographical Analysis, 8*(3), 305–313. https://doi.org/10.1111/j.1538-4632.1976.tb00538.x

Lange, R. T., & Lippa, S. M. (2017). Sensitivity and specificity should never be interpreted in isolation without consideration of other clinical utility metrics. *The Clinical Neuropsychologist, 31*(6–7), 1015–1028. https://doi.org/10.1080/13854046.2017.1335438

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist, 17*(3), 410–425. https://doi.org/10.1076/clin.17.3.410.18089

Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist, 22*(4), 666–679. https://doi.org/10.1080/13854040701494987

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. https://doi.org/10.1093/arclin/acu019

Larrabee, G. J., Greiffenstein, M. F., Greve, K. W., & Bianchini, K. J. (2007). Redefining diagnostic criteria for malingering. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits*. Oxford University Press.

Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2009). 40 Plus or minus 10, a new magical number: Reply to Russell. *The Clinical Neuropsychologist, 23*(5), 841–849. https://doi.org/10.1080/13854040902796735

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, 33*(8), 1354–1372. https://doi.org/10.1080/13854046.2019.1614227

Lawson, R. (2004). Small sample confidence intervals for the odds ratio. *Communications in Statistics - Simulation and Computation, 33*(4), 1095–1113. https://doi.org/10.1081/SAC-200040691

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. https://doi.org/10.1080/13854046.2017.1406146

Macek-Jilkova, Z., Malov, S. I., Kurma, K., Charrat, C., Decaens, T., Peretolchina, N. P., Marche, P. N., Malov, I. V., & Yushchuk, N. D. (2021). Clinical and experimental evaluation of diagnostic significance of alpha-fetoprotein and osteopontin at the early stage of hepatocellular cancer. *Bulletin of Experimental Biology and Medicine, 170*(3), 340–344. https://doi.org/10.1007/s10517-021-05063-0

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. https://doi.org/10.1080/13854046.2015.1087597

Martin, P. K., Schroeder, R. W., Olsen, D. H., Maloy, H., Boettcher, A., Ernst, N., & Okut, H. (2020). A systematic review and meta-analysis of the Test of Memory Malingering in adults: Two decades of deception detection. *The Clinical Neuropsychologist, 34*(1), 88–119. https://doi.org/10.1080/13854046.2019.1637027

Meyers, J. E., Miller, R. M., Thompson, L. M., Scalese, 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. https://doi.org/10.1093/arclin/acu001

Meyers, J. E., & Volbrecht, M. E. (2003). A validation of multiple malingering detection methods in a large clinical sample. *Archives of Clinical Neuropsychology, 18*(3), 261–276. https://doi.org/10.1016/S0887-6177(02)00136-1

Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics, 13*(3), 253–267. https://doi.org/10.1080/09720502.2010.10700699

Millis, S. R., & Kler, S. (1995). Limitations of the Rey Fifteen-Item test in the detection of malingering. *The Clinical Neuropsychologist, 9*(3), 241–244. https://doi.org/10.1080/13854049508400486

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical & Experimental Neuropsychology, 24*(8), 1094. https://doi.org/10.1076/jcen.24.8.1094.8379

Moons, K. G. M., de Groot, J. A. H., Linnet, K., Reitsma, J. B., & Bossuyt, P. M. M. (2012). Quantifying the added value of a diagnostic test or marker. *Clinical Chemistry, 58*(10), 1408–1417. https://doi.org/10.1373/clinchem.2012.182550

Myers, W. C., Hall, R., Marshall, R., Tolou-Shams, M., & Wooten, K. (2016). Frequency and detection of malingering in homicide defendants undergoing criminal responsibility evaluations using the schedule for nonadaptive and adaptive personality: A feasibility study. *SAGE Open, 6*(2), 215824401663813. https://doi.org/10.1177/2158244016638131

Ostrowski, T. R., & Ostrowski, T. (2020). The basic four measures and their derivates in dichotomous diagnostic tests. *International Journal of Clinical Biostatistics and Biometrics*, *6*(1). https://doi.org/10.23937/2469-5831/1510026

Oudman, E., Krooshof, E., van Oort, R., Lloyd, B., Wijnia, J. W., & Postma, A. (2020). Effects of Korsakoff Amnesia on performance and symptom validity testing. *Applied Neuropsychology:ADult, 27*(6), 549–557. https://doi.org/10.1080/23279095.2019.1576180

Pearson. (2009). *Advanced clinical solutions for WAIS®-IV and WMW®-IV: Clinical and interpretive manual. Pearson.*

Pepe, M. S. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Preacher, K. J. (2001). *Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software]*. http://www.quantpsy.org/chisq/chisq.htm

Proeve, M. (2009). Issues in the application of Bayes' theorem to child abuse decision making. *Child Maltreatment, 14*(1), 114–120. https://doi.org/10.1177/1077559508318395

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the test of memory malingering (TOMM). *Psychological Assessment, 10*(1), 10–20. https://doi.org/10.1037/1040-3590.10.1.10

Rey, André (1964). *L' examen clinique en psychologie* (2. éd.). Presses universitaires de France.

Rondinelli, R. D., Genovese, E., Brigham, C. R., & American Medical Association. (2008). *Guides to the evaluation of permanent impairment.* (6th ed.). American Medical Association.

Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45*(2), 294–302. https://doi.org/10.1002/1097-4679(198903)45:2%3c294::AID-JCLP2270450218%3e3.0.CO;2-N

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*(8), 851–858. https://doi.org/10.1017/S1355617716000746

Sci Stat. (n.d.). *Odd Ratio Calculator*. Retrieved September 23, 2020, from https://www.scistat.com/statisticaltests/odds_ratio.php

Shandera, A. L., Berry, D. T. R., Clark, J. A., Schipper, L. J., Graue, L. O., & Harp, J. P. (2010). Detection of malingered mental retardation. *Psychological Assessment, 22*(1), 50–56. https://doi.org/10.1037/a0016585

Sherman, E. M. S., Slick, D. J., & Iverson, G. L. (2020). Multidimensional malingering criteria for neuropsychological assessment: A 20-year update of the malingered neuropsychological dysfunction criteria. *Archives of Clinical Neuropsychology*, acaa019. https://doi.org/10.1093/arclin/acaa019

Silva, H., & Bernardino, J. (2022). Machine learning algorithms: An experimental evaluation for decision support systems. *Algorithms, 15*(4), 130–154. https://doi.org/10.3390/a15040130

Slick, D. J., & Sherman, E. M. S. (2012). Differential diagnosis of malingering and related clinical presentations. In E. M. S. Sherman & Brooks, B. L. (Eds.), *Pediatric Forensic Neuropsychology* (pp. 113–135).

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13*(4), 545–561. https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT545

Steyerberg, E. W. (2009). *Clinical prediction models: A practical approach to development, validation, and updating*. Springer.

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, Jd. F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology, 54*(8), 774–781. https://doi.org/10.1016/S0895-4356(01)00341-9

Sweet, J. J. (2009). Appendix B: Neuropsychological and psychological measures used to identify insufficient effort and malingering: A cross-referenced bibliography. In J. E. Morgan & J. J. Sweet (Eds.), *Neuropsychology of malingering casebook* (pp. 586–608). Psychology Press.

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., Kirkwood, M. W., Schroeder, R. W., & Suhr, J. A. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 35*(6), 1053–1106. https://doi.org/10.1080/13854046.2021.1896036

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (Seventh edition.). Pearson.

The Chinese University of Hong Kong. (n.d.). *C.I. Calculator: Diagnostic Statistics*. C.I. Calculator: Diagnostic Statistics. https://www2.ccrb.cuhk.edu.hk/stat/confidence%20interval/Diagnostic%20Statistic.htm#Formula

Tombaugh, T. N. (1996). *Test of memory malingering: TOMM*. Multi-Health Systems.

Underwood, J., De Francesco, D., Leech, R., Sabin, C. A., & Winston, A. (2018). Medicalising normality? Using a simulated dataset to assess the performance of different diagnostic criteria of HIV-associated cognitive impairment. *PLoS ONE, 13*(4), e0194760–e0194760. https://doi.org/10.1371/journal.pone.0194760

Victor, T. L., Boone, K. B., Serpa, K. G., Buehler, J., & Ziegler, E. A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist, 23*(2), 297–313. https://doi.org/10.1080/13854040802232682

Wang, K.-J., Makond, B., & Wang, K.-M. (2014). Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: A case study of Taiwan. *Computers in Biology and Medicine, 47*, 147–160. https://doi.org/10.1016/j.compbiomed.2014.02.002

Webb, G. I. (2017). Algorithm Evaluation. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (2nd ed., pp. 40–41). Springer US. https://doi.org/10.1007/978-0-387-30164-8_18

Witteveen, A., Nane, G. F., Vliegen, I. M. H., Siesling, S., & IJzerman, M. J. (2018). Comparison of logistic regression and Bayesian networks for risk prediction of breast cancer recurrence. *Medical Decision Making, 38*(7), 822–833. https://doi.org/10.1177/0272989X18790963

Wolfe, P. L., Millis, S. R., Hanks, R., Fichtenberg, N., Larrabee, G. J., & Sweet, J. J. (2010). Effort Indicators within the California Verbal Learning Test-II (CVLT-II). *The Clinical Neuropsychologist, 24*(1), 153–168. https://doi.org/10.1080/13854040903107791

Zadora, G., Martyna, A., Ramos, D., & Aitken, C. (2014). *Statistical analysis in forensic science: Evidential value of multivariate physicochemical data*. John Wiley & Sons Inc.

Zhou, X.-Hua., McClish, D. K., & Obuchowski, N. A. (2011). *Statistical methods in diagnostic medicine* (2nd ed.). Wiley.

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