Assessment
Volume 16 Number 3
September 2009 215-236

10.1177/1073191108325005
http://asmnt.sagepub.com
hosted at
http://online.sagepub.com

The Test Validation Summary

Richard I. Frederick
U.S. Medical Center for Federal Prisoners
Stephen C. Bowden
University of Melbourne

Common rates employed in classificatory testing are the true positive rate (TPR), false positive rate (FPR), positive predictive power (PPP), and negative predictive power (NPP). FPR and TPR are estimated from research samples representing populations to be distinguished by classificatory testing. PPP and NPP are used by clinicians to classify test takers into populations. PPP and NPP depend on the base rate (BR) of population members in the clinician's sample. The authors introduce the test validation summary (TVS) as a means to report within a single graph the FPR and TPR and the ranges of PPP and NPP across all potential sample BRs for any chosen cut score. The authors investigate how the TVS has other applications, including the estimation of local BR for the condition of interest and the estimation of standard errors for FPR and TPR when estimated across multiple independent validation studies of the classificatory test.

Keywords: *psychological testing; malingering; predictive power; test utilities*

This article concerns the use of classificatory tests. A classificatory test is developed to identify individuals who have a certain condition. To evaluate the classificatory properties of the new test, test developers select a number of individuals who have the condition, and these individuals complete the test. Typically, the researcher then evaluates a range of scores that are observed within the sample of individuals "with the condition" as potential "cut scores." A "positive" score with respect to the cut score will be used to classify an individual as "one who has the condition." Based on the scoring scheme for the test, positive scores may be either above or below the chosen cut score. The proportion of individuals in the study sample who generate positive scores is estimated to be the "true positive rate" (TPR) of the population of individuals with the condition. TPR is often referred to as "sensitivity," but we are better able to present our discussion by reference only to "positive rates." Simultaneously, the developers collect a sample of individuals who "do not have the condition" and administer the test to them. The rate at which individuals "without the condition" generate positive scores with respect to a chosen cut score is estimated to be the "false positive rate" (FPR) of the population of individuals without the condition. Most researchers report FPR by computing "specificity," which is the complement of the FPR. In other words, if the FPR = .20, the specificity = .80. Cut scores are chosen by joint consideration of FPR and TPR. The goal is to simultaneously maximize TPR and minimize FPR. Other strategies are possible, for example, finding the cut score at which FPR approximates 0, without regard to the impact on TPR.

The logic underlying the above use of classificatory tests assumes that there exists a population of individuals "with the condition" and a population of individuals "without the condition." The populations are assumed to be exclusive and exhaustive. For any test developed to distinguish population membership, the distributions of the test scores from each population overlap, such that many individuals from different populations obtain the same scores. It is unusual to find a test that perfectly separates the populations (Cohen, 1988; Woods, Weinborn, & Lovejoy, 2003).

Authors' Note: The authors thank Ross Crosby, Ira Bernstein, and Mike Speed for helpful guidance during the preparation of this article. Correspondence concerning this article should be addressed to Richard Frederick, Department of Psychology, U.S. Medical Center for Federal Prisoners, 1900 W. Sunshine St., Springfield, MO 65807; e-mail: rickfrederick@gmail.com.

FPR and TPR are conditional probabilities, representing the probability that a person from the population "with the condition" will generate a positive score. An assumption of this paradigm is that FPR and TPR are independent of any sample to which an individual belongs (Dawes, 1967). For example, if we have a test that identifies individuals who have schizophrenia, TPR represents the probability that a person with schizophrenia will generate a positive score, whether the individual is hospitalized for treatment of the condition, is in a jail, or resides in the community. There will likely be some variation in the probability of a positive test score within such subgroups of individuals with schizophrenia, and this potential variation is a matter for test researchers and test users to investigate. Again, however, when a researcher reports FPR and TPR for a test score, the general assumptions are that there are two exclusive, exhaustive populations of individuals, and the FPR and TPR are independent of sample membership. This paradigm serves as the basis of our discussions in this article.

Once the FPR and TPR are reliably estimated for the populations of interest, clinicians administer the test to an individual to obtain a score. If the score is "positive," the clinician must decide whether the positive score should be interpreted to mean that the individual has the condition. If the score is "negative," the clinician must decide whether the negative score should be interpreted to mean that the individual does not have the condition. These are the essential problems for test score interpretation by the clinician (Straus, Richardson, Glasziou, & Haynes, 2005; Woods et al., 2003).

Within samples for which classification is desirable, the FPR and TPR (*the probabilities of generating a positive test score given the respective population membership*) can be combined in some way to generate a different sort of probability: *the probability of population membership given the test score*. The latter sort of probability is referred to as "predictive power," which reflects the power of the classificatory test score to predict population membership (Elwood, 1993). Positive predictive power (PPP) refers to the power of the obtained score to predict population membership given a positive score (the probability that a positive score represents having the condition). Negative predictive power (NPP) refers to the power of the obtained score to predict population membership given a negative score. In general terminology, the TPR, FPR, PPP, and NPP values are referred to as test utilities, or test score characteristics.

A number of articles have appeared in the past 50 years, beginning with Meehl and Rosen (1955), that highlight the perils clinicians face if they indiscriminately classify members into one population when confusing TPR and FPR with PPP and NPP. Researchers work to improve estimates of TPR and FPR; clinicians must estimate PPP and NPP from these values. The two main problems that seem to prevent clinicians from deriving PPP and NPP are that (1) the calculations for predictive power often prove challenging for some and (2) the calculations require the knowledge of another value—the local base rate (BR), which may not be readily accessible (Rosenfeld, Sands, & van Gorp, 2000, Woods et al., 2003). The local BR is the proportion of individuals in the clinician's sample who have the condition.

The calculations for deriving predictive powers are,

$$PPP = TPR*BR / [(TPR*BR) + (FPR*(1 - BR))],$$

$$NPP = [(1 - FPR)*(1 - BR)]/[(1 - TPR)*BR + (1 - FPR)*(1 - BR)],$$

Our goal is to report a means to derive PPP and NPP for any given BR by observation only to obviate the need for hand calculations.

The Need to Calculate PPP for the Local BR

Because we review malingering tests in this article, our first example considers the problem of classifying a test taker as a malingerer (a member of Population A) or not (a member of Population B). In an attempt to eliminate the need for hand calculation of predictive power and to induce clinicians to use predictive power in their decision making, many test authors or researchers have supplied values for PPP and NPP in their research articles or test papers. For example, O'Bryant and Lucas (2006) reported *the* PPP for the Test of Memory Malingering (TOMM). Instead of improving the classification process, this practice of reporting fixed values of predictive powers leads to incorrect decision making because the BR of most research studies (typically close to 50% for good research design purposes) is often quite different than the local BR. Predictive power calculations should be based on local BR.

As an example, we consider a clinician who works on an inpatient mental health unit in which most of the

patients have psychotic disorders. In the past year, this clinician has evaluated all of the 99 individuals who have self-referred or who have been committed for inpatient treatment. Let us assume that all of these individuals have demonstrated bona fide psychotic illness. The 100th admission, however, presents in a strikingly unusual fashion, to the extent that the clinician begins to suspect that the presentation is feigned. She decides to administer the Structured Interview of Reported Symptoms (SIRS) as part of the evaluation process.

The SIRS (Rogers, Bagby, & Dickens, 1992) "is the single most comprehensive measure for the assessment of malingering and related response styles" (p. 6). The recommended cut score for the SIRS is "the presence of three or more scale scores in the probable feigning range" (p. 24). This criterion "correctly identified 100 of 206 (48.5%) feigners . . . with only 1 of 197 (0.5%) honest responders being misclassified" (p. 24; FPR = .005, TPR = .485). In Table 16 (p. 24), Rogers et al. (1992) reported this cut score as having PPP = .979. Based on these reported values, the PPP for the validation sample is correctly computed as 100/101, which is .99. We note that the BR of malingering for their research sample was 206/(206 + 197) = .51, so a precise statement is, "The PPP is .99 when the BR is .51."

Continuing with our example, the individual being evaluated for admission generates three probable malingering scores on the SIRS. Consequently, based on SIRS test performance and the associated reported PPP by Rogers et al. (1992) of .98, the clinician decides to classify the presentation as feigned. Is this a defensible decision based on the SIRS PPP? The decision proves *not* to be defensible because the PPP value from the SIRS manual does not apply to the clinician's circumstances. The BR of malingering in her sample is either 0% or 1% because we assumed above that the first 99 of the last 100 patients examined were genuine, and we are yet to decide about the 100th. The PPP value reported in Table 16 of the SIRS manual is based on a sample in which the BR is 51%. Because PPP depends on local BR, the PPP in her sample is certainly much lower than .98.

The SIRS has a TPR = .485 and an FPR = .005. Even if we assume that the 100th individual is actually a malingerer and therefore assume local BR = .01, the PPP for her sample is calculated as,

$$.485\ (.01) / [.485\ (.01) + .005\ (.99)] = .00485 / (.00485 + .00495) = .00485 / .0098 = .495.$$

Even though PPP = .495 seems like a toss-up, we recall that the BR of malingering up until this questionable individual had been 0. The PPP of a positive SIRS score when the BR is 0 is itself 0.

$$.485\ (0) / .485\ (0) + .005\ (1) = 0 / .005 = 0.$$

The range of potential value for the SIRS PPP given this clinician's local BR is 0 to .495. The clinician should not have depended on the value of PPP = .98 reported in the SIRS manual and should have performed the hand calculation to obtain a better estimate of PPP for her local BR.

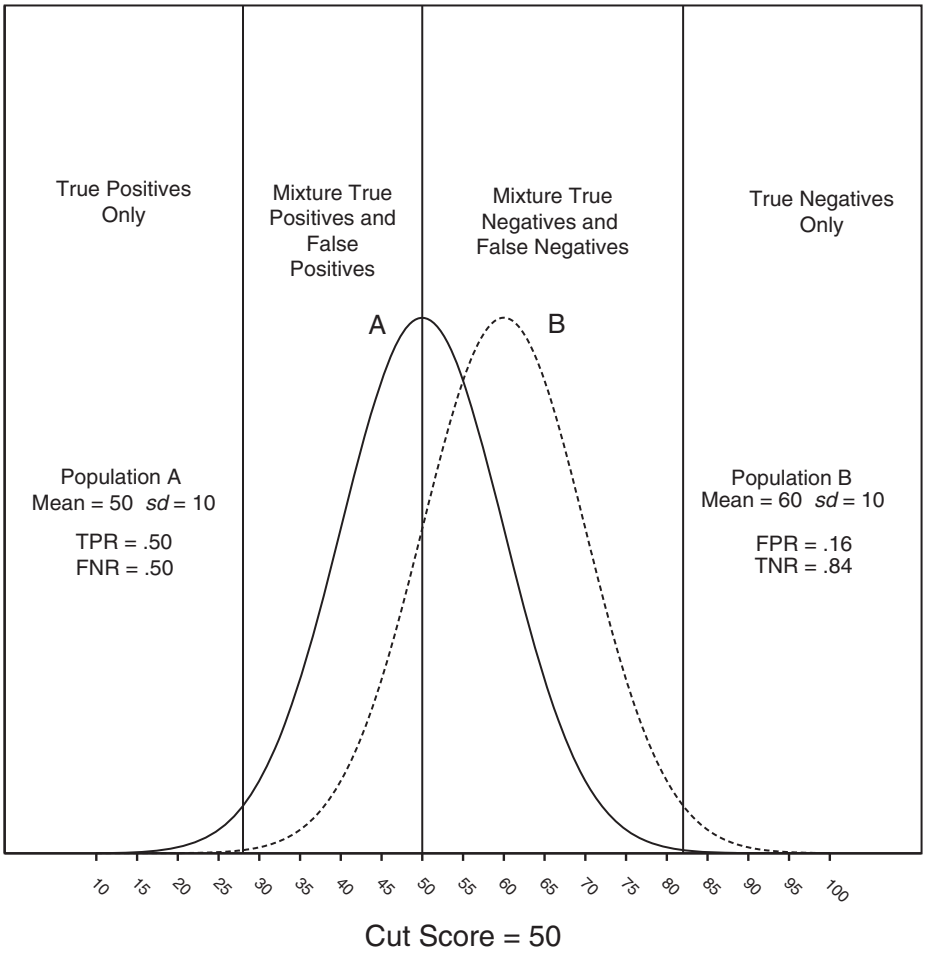
The Test Validation Summary (TVS)

The goal of this article is to introduce a method of reporting test score characteristics in a way to allow clinicians to immediately derive PPP and NPP from reliable estimates of the TPR and FPR for a cut score without hand calculations. We introduce the TVS as a graphical display of test score characteristics that, when the TPR and FPR are reliably estimated, allow a clinician to determine the PPP and NPP by immediate observation, given his or her estimate of the local BR. Because many clinicians might not be able to give a reasonable estimate of the local BR for a condition, we show how the TVS can be used to derive the local BR in situations in which a classificatory test is routinely administered.

In this article, we (a) describe the process of creating a TVS and explain its interpretation, (b) derive a TVS for the SIRS and discuss its implications, (c) identify how the TVS can help clinicians choose among all available cut scores to choose the most efficient cut score for their sample BRs, (d) identify how the TVS can help clinicians estimate their sample BRs, (e) explain how failure to consider standard errors of estimation for FPR and TPR can lead to questionable application of published test cut scores, (f) and show how the TVS can improve estimation of FPR and TPR from a number of published studies while generating stable estimates of the standard errors of estimation for FPR and TPR.

We provide a few caveats before introducing the TVS. In our discussions, individuals "with the condition" will be referred to as coming from Population A. People "without the condition" will be referred to as coming from Population B. Populations A and B exhaust all individuals with which we are concerned. Even though the combination of Populations A and B

Figure 1
Hypothetical Distributions for Two Populations of Test Scores, A and B, With Different Mean Scores (50 and 60, respectively) but the Same Standard Deviations (10)



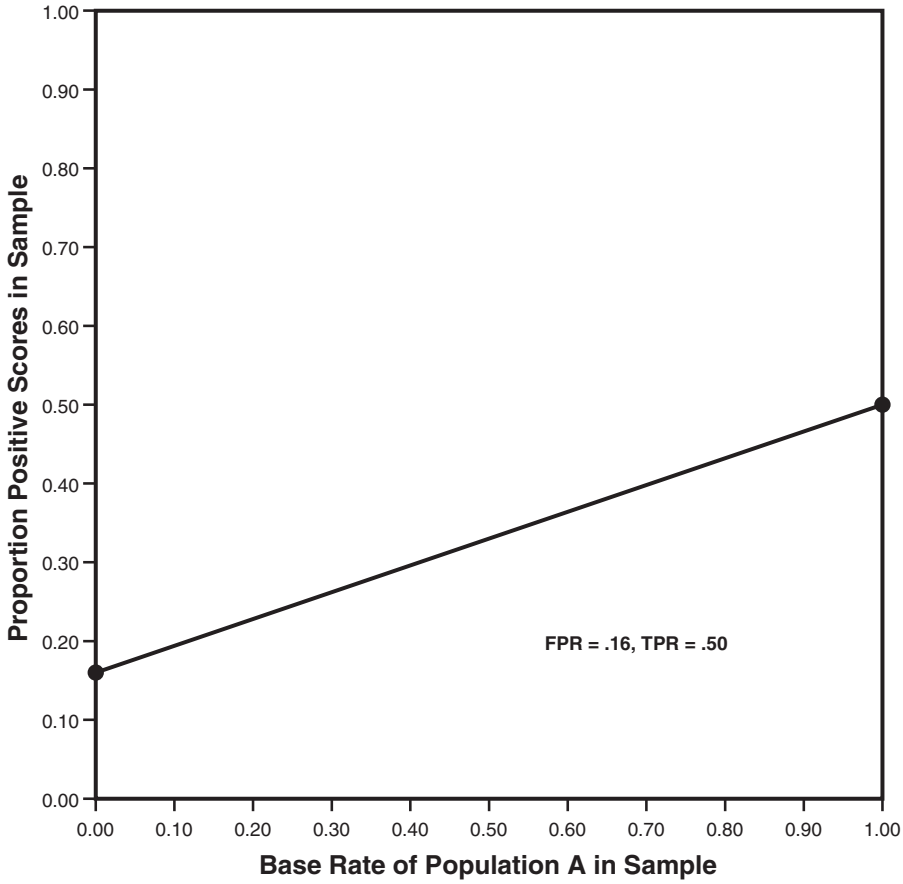
Note: TPR = true positive rate; FNR = False Negative Rate; FPR = false positive rate; TNR = True Negative Rate. The figure illustrates the classification of members of each population resulting from the application of a single decision rule or cut score of 50; that is, less than or equal to 50 is a positive score. See text for further details.

could itself be construed as a population, this does not serve our purpose. This paradigm of two exclusive, exhaustive populations serves as the foundation for the use of FPR, TPR, PPP, and NPP in classificatory testing (Baldessarini, Finklestein, & Arana, 1983; Meehl & Rosen, 1955).

Figure 1 represents the distributions of test scores for two populations, Population A (solid line) and Population B (dashed line). All individuals "with the condition" compose Population A. All individuals "without the condition" compose Population B. Members of Population A generate a mean score = 50, with $SD = 10$. Members of Population B generate a mean score = 60, with $SD = 10$. The chosen cut score is 50.

Figure 1 demonstrates the essential problem related to classifying group membership based on test score: The distributions of test scores in the populations overlap. We see that for most potential cut scores, we have mixtures of correct classifications and errors. Below the chosen cut score of 50, we have a mixture of true positive classification and false positive errors. Above the cut score of 50, we have a mixture of true negative classifications and false negative errors. When population distributions overlap to any significant degree, very few cut scores result in no false positive or false negative errors, and at those cut scores the rates of correct classifications relative to the population are generally too low to be useful.

Figure 2
The Positive Proportion Line (PPL)



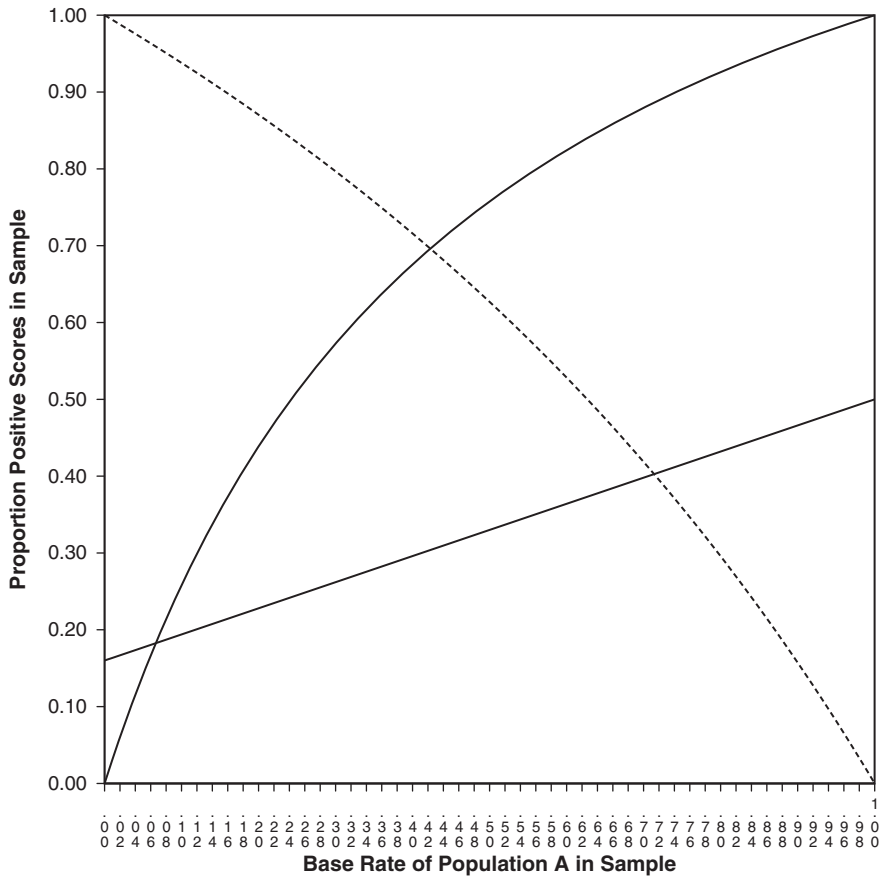
Note: Figure 2 demonstrates the range of positive test scores for a hypothetical test (i.e., a specific cut score) used to classify test takers as members of Population A or Population B. Values plotted on the x-axis reflect the rate of Population A members in the tested sample, the local base rate (BR). Values on the y-axis reflect the proportion of sample members with positive test scores. For this test, when x = 0 (when there are no members of Population A in the sample, BR = 0), y = .16 (i.e., false positive rate [FPR] = .16). When x = 1 (when there are no members of Population B, BR = 1), y = .50 (i.e., true positive rate [TPR] = .50). Connecting these two points, (0, .16) and (1, .50), plots the PPL.

Because the distributions of scores are parametric characteristics, so are the rates of classification errors at each cut score. As population distributions separate, that is, as the mean separation for a given test score between populations increases, the rates of false positive and false negative errors are minimized by choosing a cut score at the point of intersection of the distribution (Rorer & Dawes, 1982). When populations are perfectly separated, it is possible to choose a cut score that results in no classification errors. Such distributions are not commonly encountered in psychological testing (Cohen, 1988; Zakzanis, 1998, 2001).

Based on the distributions of population test scores represented in Figure 1, it is a straightforward process to determine FPR and TPR for any cut score. The cut score represented in Figure 1 is 50 or lower. From a standard z score table, we recognize that the cumulative proportion of people in Population A at score 50 or lower is 50% (TPR = .50). The cumulative proportion of people in Population B at score 50 or lower is 16% (FPR = .16).

To begin our discussion of the TVS, we construct a graph (Figure 2) in which the x-axis represents BR, with minimum value = 0 and the maximum value = 1.

Figure 3
The Test Validation Summary (TVS) for a Hypothetical Test (i.e., a Specific Cut Score) in Which False Positive Rate = .16 and True Positive Rate = .50



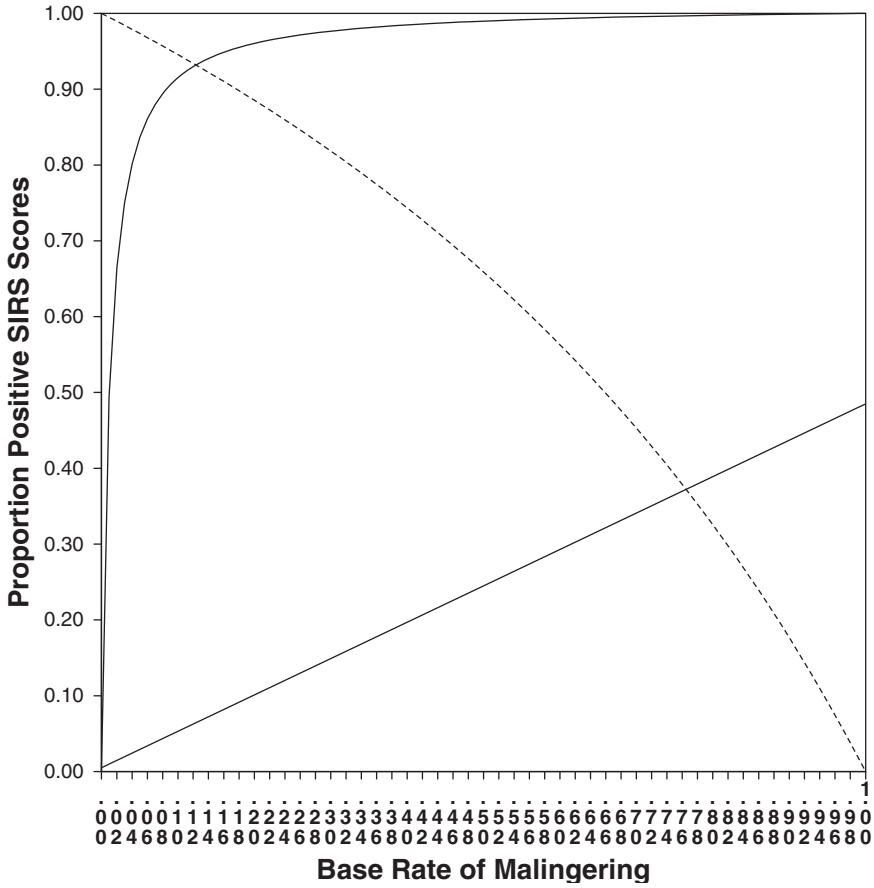
Note: The diagonal line is the positive proportion line. Positive predictive power (solid curve) and negative predictive power (dotted curve) for the test are plotted as a function of the local base rate on the x-axis. The TVS provides a succinct graphical summary of the diagnostic validity of a test across the domain of BR values. See text for further details.

We used the circumstances reflected in Figure 1 for the cut score of 50. In Figure 2, the y-axis represents the proportion of observed positive test scores (i.e., *proportion* scores ≤ 50, minimum value = 0, maximum value = 1). When there are no members of Population A in our sample, that is, when the BR or $x = 0$, the value of y is .16, which is the proportion of individuals in Population B who generate a positive score. This is the FPR of the test (plotted $x = 0$, $y = .16$). These positive scores represent test performance by Population B members only. When there are no members of Population B in our sample, that is, when $x = 1$, the value of y is .50, which is the proportion of individuals in Population A who generate a positive score. This is the TPR of the test ($x = 1$, $y = .50$). These positive scores represent test performance of Population A members only.

Thus, the effective range of y is .16 to .50. There is no reason to expect proportions of positive test scores outside the range of .16 to .50 except as a function of (a) the error associated with estimating FPR and TPR from populations and (b) the standard error of measurement associated with the test. We refer to the diagonal line that connects the (x, y) pairs ($x = 0$, $y = .16$) and ($x = 1$, $y = .50$) as the positive proportion line (PPL). (For further discussion of the derivation and meaning of this line, see Frederick [2000].) Along this diagonal line, we can determine what proportions of positive scores will be observed across the domain of BRs of Population A members in any given sample.

Figure 3 demonstrates the relationship among BR, PPP, and NPP for the illustrative test with FPR = .16 and TPR = .5 at cut score = 50. The curves in Figure 3 were derived using the formulas for PPP and NPP

Figure 4
TVS for the Structured Interview of Reported Symptoms (SIRS) Standard Decision Rule Using Reported Values of False Positive Rate = .005 and True Positive Rate = .485



Note: As revealed by the sharp differences in shapes of the positive predictive power (PPP) (solid curved line) and negative predictive power (dashed curved line), positive SIRS scores typically have greater classificatory accuracy than negative SIRS test scores. The PPP drops off rapidly as base rate of malingering (BR) drops below .10, but the PPP does not fall below .50 until the BR is as low as .01.

given above and were written into a simple spreadsheet using SPSS syntax (a copy of this syntax is presented in the appendix). The x-axis represents the BR. The straight line is the PPL. The solid curved line represents PPP; the dashed curved line represents NPP. The y-axis represents the proportion of positive scores and is also used to read the value of PPP and NPP. Although similar plots of PPP and NPP have been described before (Baldessarini et al., 1983; Straus et al., 2005), the nonlinear function relating the BR to both PPP and NPP is not well appreciated among clinicians, as indicated by the number of published reports of fixed values of PPP and NPP.

When BR = 0, the PPP is 0 and NPP is 1. That is, when no one has the condition, a positive test score means the best classification is "does not have the condition." When BR = 1, the PPP is 1 and NPP is 0. That is, when everyone has the condition, a positive test score means the best classification is "has the condition." This graph immediately communicates the relationship among BR, the proportion of positive test scores, PPP, and NPP. We see that the PPP does not attain a value of 50% accuracy until the BR is about 26%, so it may not make sense to use this hypothetical test to identify people "with the condition" in samples with lower than 26% BR.

To present this discussion most easily, we have represented a normal distribution of test scores in both populations. Test score distributions of interest might not be normal. The distribution shape can best be estimated by reporting TPR and FPR for the entire set of cut scores for a test, as in receiver operating

characteristic curve analysis. Estimating the distribution shape in this way may motivate the choice of a cut score. Once a cut score is justified and TPR and FPR are reported for that cut score, the TVS is useful no matter what shape the parent distribution takes, whether normal or not. If the TPR and FPR are justifiably used, the TVS effectively demonstrates PPP and NPP.

SIRS TVS

Figure 4 is a TVS for the SIRS using the FPR (.005) and TPR (.485) for the standard decision rule for malingering reported in the SIRS manual (Rogers et al., 1992, p. 24; i.e., one "definite" or three or more "probables"). We note the end points of the PPL are at $y = .005$ (the FPR of the SIRS) and at $y = .485$ (the TPR of the SIRS). Because the SIRS has an extremely low FPR, the PPP (solid curved line) for the SIRS exceeds .90 for most BRs > .10. Consistent with our previous example of a clinician with one suspicious inpatient mental health admission, we see that when the BR of malingering is .01, the PPP is estimated at about .50.

The TPR of the SIRS reported in the manual is .485. That is, more than half the participants who were malingering in the SIRS validation sample received a negative SIRS score. That might suggest to some that when a negative score is obtained, there might be a high probability that it results from malingering. But we note in the SIRS TVS that the NPP is also rather high when the BR of malingering is near 10%. When the BR = .10, the likelihood that a negative score represents "not malingering" is greater than .90. The SIRS TVS greatly clarifies decision making when using the SIRS in samples with BRs far lower than the validation sample BR of .51.

Considering Different Cut Scores for Clinical Classification

Clinicians may wish to use a different cut score if the obtained score of the person to be classified does not result in a high enough PPP or NPP to afford reliable classification. Some test manuals or research articles report FPR and TPR for more than one cut score. In those circumstances, the clinician may wish to construct and consult different TVSs to evaluate alternative cut scores.

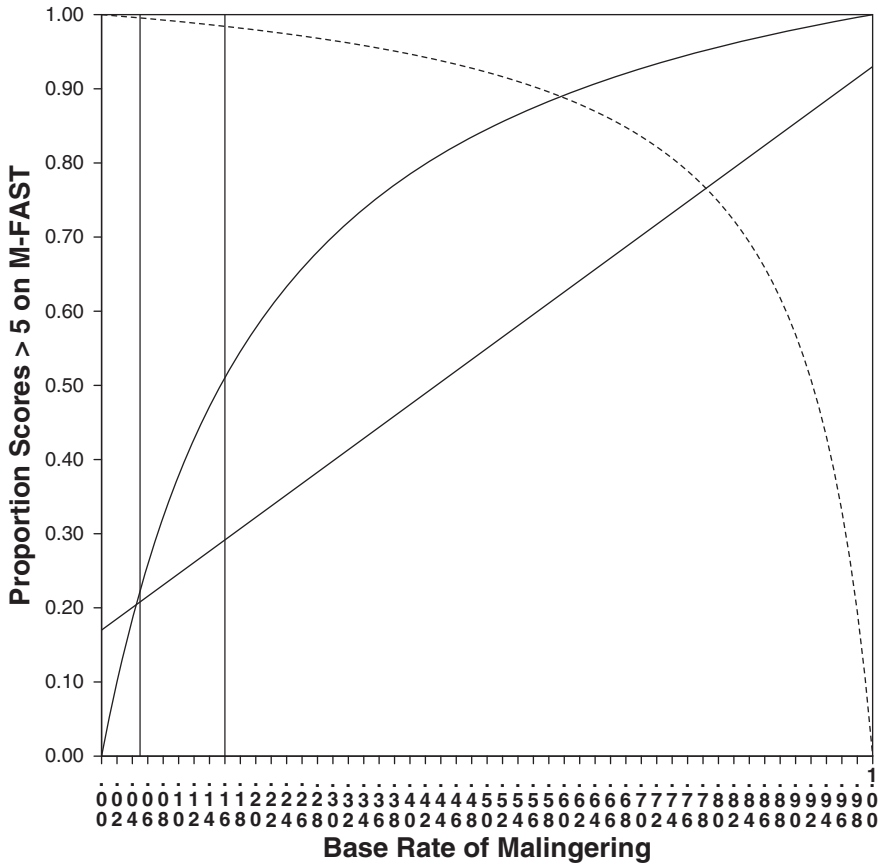
Miller Forensic Assessment of Symptoms Test (M-FAST)

The M-FAST (Miller, 2001) "is a structured interview designed to provide information regarding the probability that an individual is malingering psychiatric illness" (p. 3). Potential scores on the M-FAST range from 0 to 25. The recommended cut score for the M-FAST is 6 or higher and "was chosen to maximize Negative Predictive Power [i.e., the probability that a score of 5 or lower represents cooperation] without decreasing the Positive Predictive Power to any considerable extent" (p. 11). In her Table 4, Miller (2001, p. 12) reported the following test characteristics for the recommended cut score: Score ≥ 6, TPR = .93, FPR = .17, PPP = .68, NPP = .97. Miller's Table 4 is based on 86 individuals in a "clinical setting," with BR of malingering reported as .35. In light of our caution above, we note that this table is one in which the NPP and PPP estimates are accurate only for a single BR, when the BR of malingered psychopathology is .35.

Figure 5 is the TVS for the M-FAST values reported for a cut score of 6 (i.e., score ≥ 6) in a clinical sample. We note that the PPP of a cut score of 6 is .67 only when the BR = .35, as reported in the test manual. PPP varies from 0 to 1 as a function of BR. Unlike the obscure calculations that are required to make use of the information in Miller's (2001, p. 12) Table 3 when the local BR is not .35, the M-FAST TVS (for cut score ≥ 6) quickly communicates estimates of PPP for test users who have reasonable estimates of their local BR of malingered psychiatric symptoms. For example, when the M-FAST is used in a clinic that has a BR of malingered psychiatric symptoms of 5%, a cut score of 6 on the M-FAST has PPP = .22 (see vertical line at BR = .05 in Figure 5). Therefore, use of the recommended cut score when a clinic's BR of malingering is .05 will lead to a incorrect conclusion that someone is malingering 78% of the time. A cut score of 6 does not even approach 50% certainty about malingering until the local BR of malingered psychiatric symptoms is .16 (i.e., PPP > .50 only when BR ≥ .16; see vertical line at BR = .16 in Figure 5).

One way to approach the problem of PPP less than .50 when the BR is low is to modify the cut score so that the FPR is reduced. That is, the PPP can be improved by adopting a more conservative cut score that identifies fewer false positives. FPR and TPR at different cut scores must be supplied in the test validation study for this strategy to be feasible. For the M-FAST, Miller reported FPR and TPR at every cut score.

Figure 5
Test Validation Summary (TVS) for the Miller Forensic Assessment of Symptoms Test (M-FAST) Using Reported False Positive Rate = .17, true positive rate = .93 for the Recommended Cut Score to Determine the Positive Proportion Line



Note: The TVS is completed by plotting positive predictive power (PPP; solid curved line) and negative predictive power (dashed curved line) for all base rates (BRs) in which the test can be administered. Two vertical lines identify the PPP values for BR = .05 (PPP = .22) and BR = .16 (PPP = .50). This means the recommended M-FAST cut score yields more correct than incorrect decisions only when BR ≥ .16.

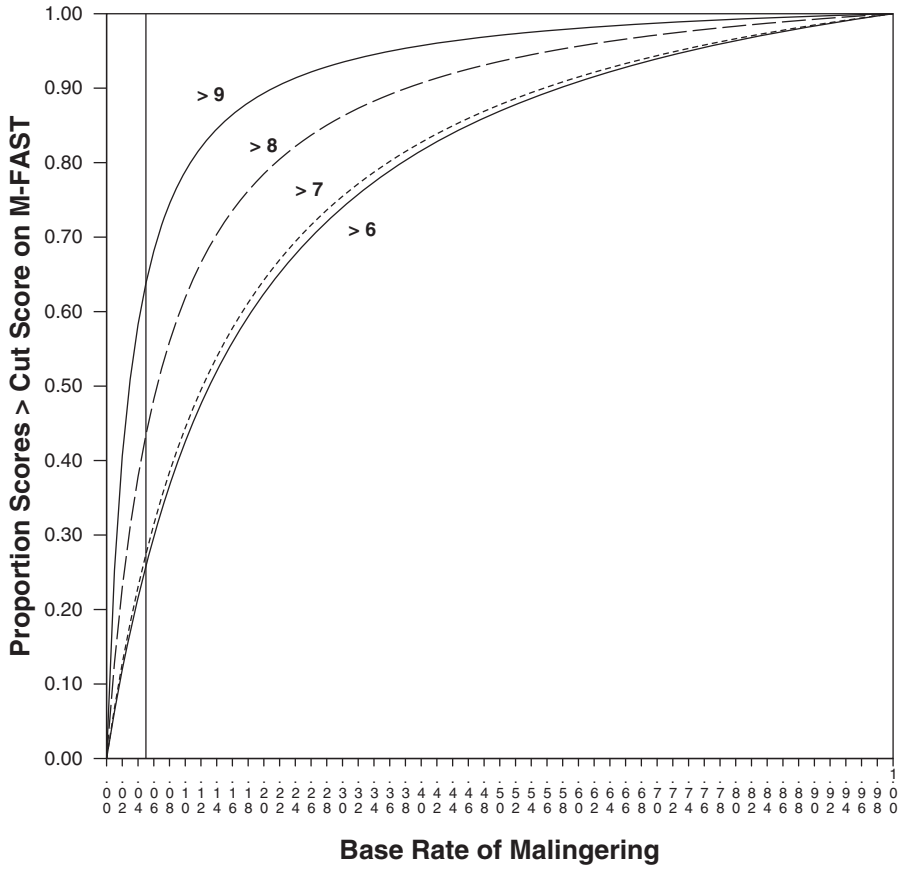
Figure 6 is a partial TVS with many PPP curves generated for cut scores greater than or equal to 7, 8, 9, and 10. Using the values reported in Miller's Table 3 (2001, p.12), we note we will not exceed 50% accuracy for identifying malingering (PPP) at BR = .05 until we employ a cut score greater than or equal to 10 (at BR = .05, for cut score > 9, PPP = .64).

Determining the Local BR From the TVS

One advantage of the TVS is that, when the FPR and TPR of a given cut score are reliably estimated, knowing the local BR allows for prediction of the proportion of positive scores that will be observed. The PPL is ultimately the intersection of the observed proportion of positive scores for each potential BR. Conversely, then, if one knows the proportion of positive scores in a sample, the local BR is determined by the PPL. That is, any *y* value along the PPL represents proportions of positive test scores within any sample that may contain a mixture of members of Populations A and B. The corresponding *x* value at this point estimates the BR of Population A members in our mixed sample.

In Figure 7, we refer back to the hypothetical test represented in Figure 2 with FPR = .16 and TPR = .50. Figure 7 supposes that 33% of a sample of individuals in a mixed sample obtained a positive score.

Figure 6
Test Validation Summary for the Miller Forensic Assessment of Symptoms Test (M-FAST) Using Cut Scores of 7 or Higher to 10 or Higher (Reported as > 6 to > 9 in the Figure)



Note: According to Miller (2001, Table 3, p. 12), the false positive rate (FPR) and true positive rate (TPR) for these cut scores are (> 6, FPR = .14, TPR = .93; > 7, FPR = .11, TPR = .79; > 8, FPR = .05, TPR = .73; > 9, FPR = .02, TPR = .67). When base rate (BR) = .05 (at vertical line), the positive predictive power (PPP) for each cut score is, respectively, from > 6 to > 9, .26, .27, .43, and .64. Thus, the PPP does not exceed 50% at BR = .05 until the cut score is raised to 10 or higher (> 9). The recommended cut score by the test author is 6 or higher, which has PPP = .22 at BR = .05.

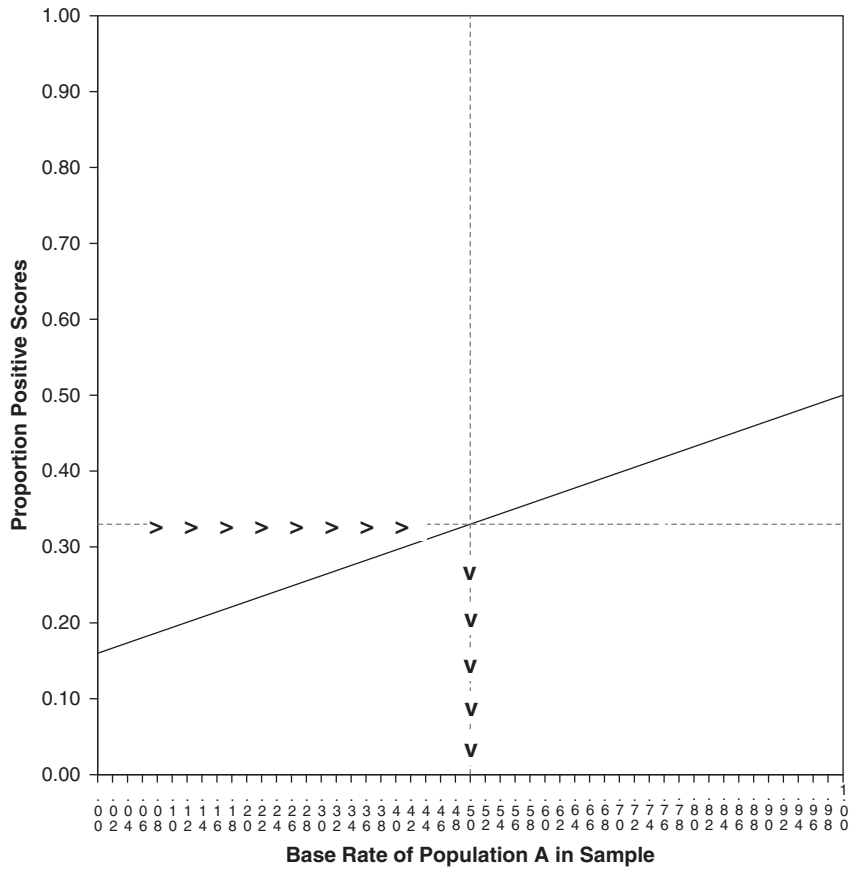
At $y = .33$ on the PPL, the x value = .50. So our estimate of local BR of Population A members is .50 when the proportion of positive scores is .33.

Using the SIRS to Estimate BR in Research Groups

Using the TVS to determine the BR in a sample of individuals who have completed a classificatory test has implications for developing research groups. For example, Rogers, Tillbrook, and Sewell (2004) examined the Evaluation of Competency to Stand Trial–Revised (ECST-R), "a standardized interview for assessing the underlying dimensions of [competency to stand trial] . . . [and] designed to screen systematically for feigned incompetency" (Rogers, Jackson, Sewell, & Harrison, 2004, p. 139). In this study, performance on the SIRS was used to form criterion groups, which the authors referred to as "known" groups.

Rogers, Jackson, et al. (2004) established validation groups by administering the SIRS to 56 individuals participating in competency restoration groups in a state hospital. Eight of these individuals generated positive SIRS cut scores, and 48 did not. Two groups were formed from the SIRS cut score, a "probable fake" group ($n = 8$) and a genuine, "clinical," group ($n = 48$). Therefore, the rate of positive test scores

Figure 7
Same Information as Figure 1 for the Hypothetical Test



Note: Figure 7 hypothesizes a sample of test takers in which 33% of test scores are positive. By plotting this value (.33) on the y-axis, one can locate the corresponding *x* value on the positive proportion line. Here, the *x* value = .50. This value represents the best estimate of the BR of Population A members in the new sample, given the reported values of false positive rate and true positive rate.

was 8/56, or .14. From the PPL in the SIRS TVS (Figure 4), we observe that a rate of 14% positive scores corresponds not to BR = .14 but to BR = .28. We can then locate the corresponding PPP at BR = .28 to estimate PPP = .975 and NPP = .83.

Given the estimated BR = .28, of the 56 individuals in the Rogers, Jackson, et al. (2004) study, the best estimate is that 16 (or 28% of 56) were malingering, but only 8 of the 16 malingerers received a positive SIRS score. This makes sense because the SIRS TPR = .485. That is, when the test is administered to a new clinical sample comprising a mixture of genuine and malingering patients, we expect only 48.5% of the malingerers to obtain a positive test score. The estimated 8 malingerers who did not receive a positive score were therefore most likely assigned to the "clinical" group. Within the group classified as "clinical" (i.e., "not malingering"), the rate of malingering therefore must have been about 8/48, or .17. This corresponds with NPP = .83, which estimates that 83% of negative scores at BR = .29 represent "not malingering."

This has strong implications for the validation of the ECST-R. A stated intention of the Rogers, Jackson, et al. (2004) study was a maximization of "sensitivity [TPR] and especially NPP" in the development of cut scores for ECST-R malingering scales (p. 143). We note that the apparent underestimation of the rate of malingering in the "clinical" group inevitably results in higher estimates of FPR because the genuine "clinical" group contains more malingerers than was thought to be the case. Consequently, when malingerers inadvertently assigned to the "clinical" validation group produce positive ECST-R malingering test scores, the computed FPR is necessarily inflated. This inflation of

FPR is misattributed to test characteristics and not to the impurity of the "known" clinical group. Frederick (2000), following Dawes and Meehl (1966), offered a solution (mixed group validation) for more accurate determination of TPR and FPR when validation groups are "mixed" in this way and not "pure." Mixed group validation is easily accommodated by the TVS, but an exploration of this application is outside the scope of this article.

Attending to Standard Errors of Estimation for FPR and TPR

Classificatory testing ultimately concerns identifying to *which* population an individual belongs. When test authors or researchers estimate TPR and FPR, they are estimating the rates at which members of each population generate scores that will be deemed positive by the chosen cut score. Therefore, the TPR and FPR are *parameters* of Populations A and B. The parametric values of TPR and FPR are static characteristics of the *population*. Sampling will produce variable estimates, and the standard error of estimation is influenced by how many individuals are sampled in research.

To this point, we have shown that the TVS rapidly communicates the values of PPP and NPP, the clinician's classification tools, at every potential local BR when the FPR and TPR of a test have been reliably estimated. We have shown how a clinician or researcher can estimate the BR of people "with the condition" in their samples, and we have shown how researchers might use that information when validating new tests. We next consider the problem test users face in choosing among reported estimates for FPR and TPR to use.

To estimate population classification rates, rioesearch samples must be representative of the populations for which the test is intended. Despite new standards to report confidence intervals for parametric estimates (Wilkinson & Task Force on Statistical Inference, American Psychological Association, Science Directorate, Washington, DC, 1999), confidence intervals are rarely reported for estimates of FPR and TPR. We consider a test in which overt consideration of sample representativeness and construction of confidence intervals based on standard errors of estimation would have been instructive to potential users of the Structured Inventory of Malingered Symptomatology (SIMS).

SIMS

The SIMS (Widows & Smith, 2005) is a "75-item, multi-axial, self-administered screening measure used for the detection of malingering across a variety of clinical and forensic settings" (p. 1). The "final version of the SIMS was empirically evaluated at four different universities and community colleges in an analogue simulation study . . . with a nonclinical sample of 476 undergraduate students who volunteered in exchange for extra class credit" (p. 12). In their Table 4.2 (Widows & Smith, 2005, p. 13), the authors reported TPR = .95 and FPR = .12 (n = 238) for the final SIMS cut score (Total Score > 14). In support of these reported rates, the authors cited Edens, Otto, and Dwyer (1999), who reported TPR = .96 and FPR = .09 for Total Score > 14 for a sample of 196 college students who first took the test honestly and then simulated a variety of modes of faking. Finally, the authors cited Merckelbach and Smith (2003), who evaluated 241 honest normals and 57 simulating malingerers, and reported for a cut score of Total Score > 16, TPR = .93 and FPR = .02. Consistent with an increase in the cut score from 14 to 16, a reduced number of honest normals were misclassified, yielding a lower FPR. Together, these three studies show consistency in high values for TPR and low values for FPR for the two investigated cut scores.

One question for a user of this test is, "What are the FPR values for these SIMS cut scores SIMS in the samples I will test?" The answer is not necessarily provided by the previous three studies, which did not include clinical patients. Only if groups of "honest" clinical patients and "honest normal" research participants both respond to the SIMS in the same way will the FPR be of the same magnitude. Therefore, it is important to determine if "honest normal" participants are representative samples of the population of clinical patients who do not appear to be malingering.

Lewis, Simcox, and Berry (2002) evaluated 55 criminal defendants undergoing mental health evaluations with the SIMS. Based on clinical examination of the defendants, two groups of 31 "honests" and 24 "malingerers" were established. For these two groups, using a cut score of Total Score > 16, the TPR was 1, and the FPR was .39. The FPR estimate appears to be substantially different from values obtained in each of the three prior studies.

To evaluate whether the Lewis et al. (2002) estimate is statistically different from the estimates

reported by Widows and Smith (2005), one can construct a confidence interval around the derived FPR of .39. The *SE* for a single proportion (*p*) is given by (Fleiss, 1981, p. 13). For $p = .39$, with $n = 24$, the *SE* is computed as .10. A 95% confidence interval for the FPR = .39 +/– (1.96*.10), which is .194 to .586; the 99% confidence interval for FPR = .39 +/– (2.58*.10) = .132 to .648. No FPR estimate for any of the three simulation studies reported for the SIMS in Widows and Smith (i.e., .12, .09, and .02) is within the 99% confidence interval generated by the Lewis et al. data.

The Merckelbach and Smith (2003) study and the Lewis et al. (2002) study both based estimates of FPR on the same cut score, Total Score > 16. Merckelbach and Smith reported FPR = .02, $n = 241$. This FPR value is not in the 99% confidence interval (i.e., .132 to .648) derived from Lewis et al. The *SE* of the FPR reported by Merckelbach and Smith is computed as .01. The 99% confidence interval for their reported FPR is 0 to .056. There is no overlap in the 99% confidence intervals derived from the Merkelbach and Smith (0 to .056) and the Lewis et al. (.132 to .648) studies, which strongly supports a conclusion that clinical patients and "compliant" students belong to two different populations of SIMS test scores. Consequently, the three studies reported by Widows and Smith (2005) have little bearing on our understanding of the potential for false positive errors in using the SIMS as a clinical tool. Users of the SIMS should probably employ a cut score much higher than 16 (which is already higher than the authors' recommended cut score) if they wish to avoid a substantially high rate of false positive errors. Because higher cut scores will substantially lower not only the FPR but also the TPR, use of the SIMS may not be practical.

Using the TVS to Estimate FPR and TPR From Multiple Studies

In our final example, we show how the TVS can facilitate estimation of FPR and TPR using other estimates of FPR and TPR from many published studies. This is possible when a large number of data points (x = BR, y = proportion positive scores) are used to generate a line of best fit, which can be construed as the PPL of a TVS. The end points of the PPL will estimate the FPR and the TPR. Furthermore, it is possible to generate *SE*s in this process, so that confidence intervals for our estimates can be generated, to be compared against estimates in future research.

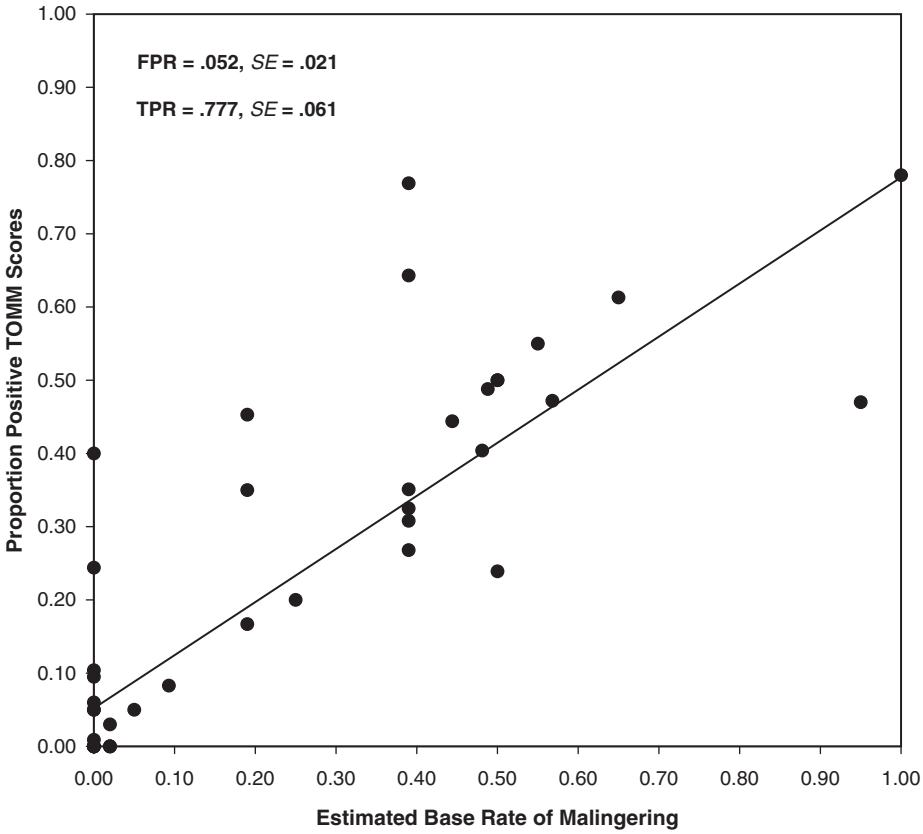
TOMM

We chose for our example the TOMM (Tombaugh, 1997) because many studies have reported the diagnostic properties of this test. The TOMM is "a 50-item recognition test for adults that includes two learning trials and a retention trial" (p. 1). A failure on the TOMM is represented by performing at less than 90% correct recognition on Trial 2 or on the Retention Trial. According to Tombaugh (1997), "While the TOMM is sensitive to malingering, it is insensitive to neurological impairments" (p. 1). Although the FPR rate of the TOMM is generally considered to be low (e.g., Rees, Tombaugh, Gansler, & Moczynski, 1998), the TPR has been estimated by some as also quite low (Gervais, Rohling, Green, & Ford, 2004; Green, 2007). To evaluate TOMM FPR and TPR, we constructed a TOMM TVS (Figure 8) by using as ordered pairs (BR, rate positive scores at the recommended cut score) for data derived or estimated from 40 data sets in 25 published articles that included the TOMM. The 40 ordered pairs were subjected to weighted least squares (WLS) regression (the n in each study was the weighting variable) to generate the end points of the PPL (i.e., FPR and TPR), which were then used as the basis to generate the PPP and NPP curves.

Table 1 summarizes the data used for TVS construction. Some articles reported more than one study; "study" represents each independent study reported within an individual article. Table 1 identifies what type and number of participants were included in the studies (college students, normal adults, children, neuropsychology examinees, criminal defendants, or psychiatric patients). The next table entry is the BR estimate across all samples used in a single study. If authors provided a means of estimating the BR of malingering in their samples, we have followed that, even if their estimates seemed faulty to us.

When no sample BR estimate appeared in the study, we followed Mittenberg, Patton, Canyock, and Condit (2002) and used a rate of .39 for litigating neuropsychology cases and .19 for criminal defendants, based on their large-scale, national sample of neuropsychological evaluations in a variety of contexts. Although it might be objected that using the Mittenberg et al. estimates is ad hoc, our view is common research practices approach capriciousness. Namely, the widely used approach of assigning BR values of 0 or 1 in criterion groups, or *known* groups validation (as promulgated by Rogers, Jackson, et al., [2004] or by Greve, Bianchini, Love,

Figure 8
Positive Proportion Line for the Test of Memory Malingering (TOMM; Tombaugh, 1997) at Recommended Cut Score



Note: FPR = false positive rate; TPR = true positive rate. In this figure, the positive proportion line is the best-fitting weighted least squares regression line derived from proportions of positive test scores (failures) regressed on estimated base rates for all 40 studies in Table 1.

Brennan, and Heinly [2006]), assumes that groups of malingerers and controls have pure composition. This assumption is obviously unfounded in most cases (Rohling & Boone, 2007). Instead, use of independently developed estimates of BRs such as those published by Mittenberg et al. more likely approximate the true values than assignments of 0 or 1. We freely acknowledge that the Mittenberg et al. values may not accurately represent the true BRs of the studies we included, and we identify in our discussion reasons to believe that the values are not precisely on target but not improbable. Our use of Mittenberg et al. employs estimates developed independently of our work, and our ultimate goal is to encourage researchers to better estimate the BR of their samples (e.g., as we described in an earlier section of this article) so that a mixed groups methodology can be employed to better estimate FPRs and TPRs of the test under validation.

Greve, Bianchini, Love, et al. (2006) and Greve, Bianchini, and Doane (2006), for example, have been evaluating how to refine estimates of the likelihood of feigning in clinical samples. By application of the "Slick criteria" (Slick, Sherman, & Iverson, 1999), they have been sorting clusters of patients with rubrics such as "No Incentive," "Incentive Only," "Suspect," "Statistically Likely," "Probably Malingering," and "Definitely Malingering." They have not assigned or derived mean probabilities for these ordinal categories. Because their criterion groups validation of test scores was inherently limited to instances in which BR = 0 or 1, they nevertheless were required to ignore the BR differences they had identified in clusters that obviously did not have BRs of 0 (e.g., Incentive Only) or 1

Table 1
Studies Used in Mixed Groups Validation (MGV) Retrospective Analysis of Test of Memory Malingering (TOMM)

Study	Participants	*N*	BR	FR	Type
Ashendorf, Constantinou, and McCaffrey (2004)	Community older adults	197	.000	.000	Normals
Bolan, Foster, Schmand, and Bolan (2002)	Community-dwelling adults	32	.500	.500	Simulators
Bolan et al. (2002)	College students	40	.500	.500	Simulators
Constantinou and McCaffrey (2003)	Normal children 5 to 12	128	.000	.000	Challenge
Cragar, Berry, Fakhoury, Cibula, and Schmitt (2006)	Seizure patients	80	.000	.050	Neuropsych
Delain, Stafford, and Ben-Porath (2003)	Criminal defendants	64	.190	.453	Crim. def.
Donders (2005)	Children patients	100	.020	.030	Challenge
Duncan (2005)	Psychotic	50	.000	.060	Challenge
Etherton, Bianchini, Heinly, and Greve (2006)	Pain	60	.000	.000	Challenge
Gavett, O'Bryant, Fisher, and McCaffrey (2005)	TBI litigants	77	.390	.325	Neuropsych
Gierok, Dickson, and Cole (2005)	Psychiatric	20	.000	.050	Challenge
Gierok, et al. (2005)	Criminal defendants	20	.190	.350	Crim. def.
Greve, Bianchini, and Doane (2006)	TBI no incentive	14	.020	.000	Neuropsych
Greve, et al. (2006)	TBI definite	9	1.00	.780	Neuropsych
Haber and Fichtenberg (2006)	TBI nonlitigants	18	.020	.000	Neuropsych
Haber and Fichtenberg (2006)	TBI litigants	28	.390	.643	Neuropsych
Hill, Ryan, Kennedy, and Malamut (2003)	Epileptic patients	48	.000	.104	Neuropsych
McCaffrey, O'Bryant, and Fisher (2003)	TBI litigants	97	.390	.268	Neuropsych
Moore and Donders (2004)	TBI mix comp noncomp	132	.093	.083	Neuropsych
Moss, Jones, Fokias, and Quinn (2003)	TBI comp seeking	102	.390	.308	Neuropsych
O'Bryant, Hilsabeck, Fisher, and McCaffrey (2003)	TBI litigants	94	.390	.351	Neuropsych
Powell, Gfeller, Hendricks, and Sharland (2004)	College students	80	.650	.613	Simulators
Rees, Tombaugh, Gansler, and Moczynski (1998)	TBI litigants	13	.390	.769	Neuropsych
Rees et al. (1998)	TBI nonlitigants	13	.020	.000	Neuropsych
Rees et al. (1998)	College students	20	.550	.550	Simulators
Rees et al. (1998)	College students	44	.568	.472	Simulators
Rees et al. (1998)	College students	18	.444	.444	Simulators
Rees, Tombaugh, and Boulay (2001)	Depressed inpatients	26	.000	.000	Challenge
Tan, Slick, Strauss, and Hultsch (2002)	College students	52	.481	.404	Simulators
Teichner and Wagner (2004)	Memory complaints	78	.000	.244	Neuropsych
Tombaugh (1997)	Community dwelling adults	405	.000	.009	Normals
Tombaugh (1997)	Normal adults	70	.000	.000	Normals
Tombaugh (1997)	Neurologically impaired	158	.000	.095	Neuropsych
Tombaugh (1997)	College students	41	.488	.488	Simulators
Vickery et al. (2004)	Patients and normals	92	.500	.239	Simulators
Weinborn, Orr, Woods, Conover, and Feix (2003)	NGRI civil committed	36	.000	.400	Challenge
Weinborn et al. (2003)	Criminal defendants	25	.190	.167	Crim. def.

Note: BR = base rate; FR = fail rate; TBI = traumatic brain injury; NGRI = not guilty by reason of insanity. BR was determined by collapsing validation groups into one group. Number of members of "malingering" validation group was divided by total to derive BR. When BR was not available from study description, .390 was used for TBI litigants and .19 was used for criminal defendants, following Mittenberg, Patton, Canyock, and Condit (2002). FR on TOMM is determined by proportion participants who scored less than 90% correct on Trial 2 or Retention Trial. For type, challenge means the participant sample represented conditions or severity of impairment arguably more likely to produce failure than the validation sample of the TOMM.

(Probably Malingering), and they treated these research groups as if they did have BRs of 0 or 1—or they were required to eliminate middle-ground research groups (e.g., Suspect or Statistically Likely to be malingering) from their estimations of FPR and TPR for test scores under consideration. For our analysis of the TOMM, we have selected from the Greve, Bianchini, and Doane (2006) study only the "No Incentive to Malinger" group, which we assigned a BR = 0, and the "Definitely Malingering" group, which we assigned a BR = 1. Had they assigned BR estimates to the other groups, we would have included them in our analysis because our analysis is not restricted to choices of 0 or 1.

In Table 1, we also report the failure rates on the TOMM. Researchers employed a variety of means to report results from the TOMM. For example, some researchers report Trial 2 failures and Retention Trial failures separately, with no overall rate of failure reported. We have read the reports carefully to identify the overall failure rates reported in Table 1, but it is possible that we have made slight mistakes in determining some failure rates. Finally, in Table 1, we characterized study types as normal, patient, simulator, or challenge. "Normal" means that the study used normal adults; these studies are considered to have malingering BR = 0, unless otherwise estimated by the authors. "Patient" means the study used litigating or nonlitigating neuropsychology examinees or mental health patients. "Simulator" means the study included normal individuals with instructions to pretend to be impaired when taking the test, even though not all participants may have been asked to simulate malingering. "Challenge" means the study used people from populations not validated for the TOMM to determine characteristics of TOMM performance.

Proportions of positive scores (failure rates) were regressed on estimated BRs in all studies reported in Table 1. In Figure 8, we derived the PPL as the line of best fit generated by WLS regression. The best fitting regression line produces *y* intercepts that yield the estimates of classificatory accuracy: FPR = .048 (*SE* = .021) and TPR = .825 (*SE* = .065).

To better estimate clinical conditions, Figure 9 shows the PPL when simulator and challenge data have been excluded from the analysis, leaving 24 studies to provide data points for WLS regression. Elimination of simulator and challenge data resulted in estimates of FPR = .049 (*SE* = .026) and TPR = .855 (*SE* = .111). These estimates are not significantly different from those obtained from the full data set in terms of the 95% confidence intervals.

Both TVSs allow us to quickly identify potential problem points in the data. In both Figure 8 and Figure 9, the problem points are at BR = .19 and BR = .39 (points of our uncritical use of Mittenberg et al. [2002] estimates). We suspect the problematic values in these ordered pairs are the BR estimates. Thus, the TVS can be used to identify instances in which we might reconsider those estimates.

We also note the number of points at BR = 0 in Figure 9 for which the rate of positive test scores exceeds 0. These are instances in which the TOMM was administered to people with significant obvious impairment (e.g., people with significant dementia), and these sorts of people are not likely in need of motivational assessment. We might have classified those participants as "challenge," except that the TOMM validation studies included these types of patients.

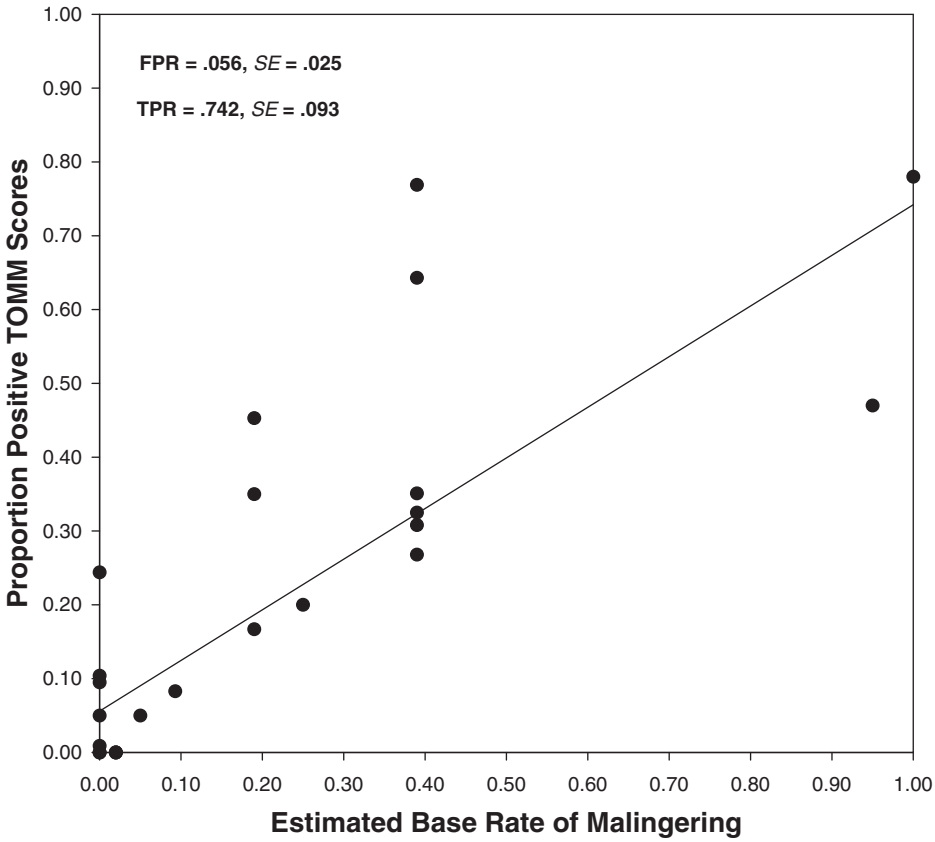
We further note that only two points in our data are for BR > .50. The TPR (the intercept at BR = 1) potentially has more error associated with its estimation than the FPR (the intercept at BR = 0) because of this paucity of points used for estimation. This is reflected in the difference in *SE* estimates for TPR (.093) as compared to FPR (.025). Estimation of predicted values in regression is more error prone at the ends of the regression line than at middle values. In our example, however, it is clear that the middle values are more unreliably estimated than the points at low BR. We continue to emphasize that research must focus on improving estimation of BR in research samples.

Figure 10 is a TVS based on FPR and TPR estimates from Figure 9. Having constructed the TOMM TVS from a large number of independent clinical studies ($n = 24$) without simulators or challenge groups, we can safely conclude that the TOMM is highly sensitive to intention to perform poorly, which contradicts the assertions made by Gervais et al. (2004) and Green (2007) that the TOMM has a low TPR.

Discussion

In this article, we have asserted that the most important classification characteristics of a test are FPR and TPR. Estimating FPR and TPR across a range of cut scores is the only way to derive the nature of population distributions of test scores for diagnostic and classification purposes. Because estimating FPR and TPR is the process of estimating a parametric property of test scores for certain populations

Figure 9
Positive Proportion Line for the Test of Memory Malingering (TOMM) Representing the Best-Fitting Weighted Least Squares Regression Line, Using Data From Studies Reported in Table 1 but Excluding Studies of Simulators and Challenge Groups



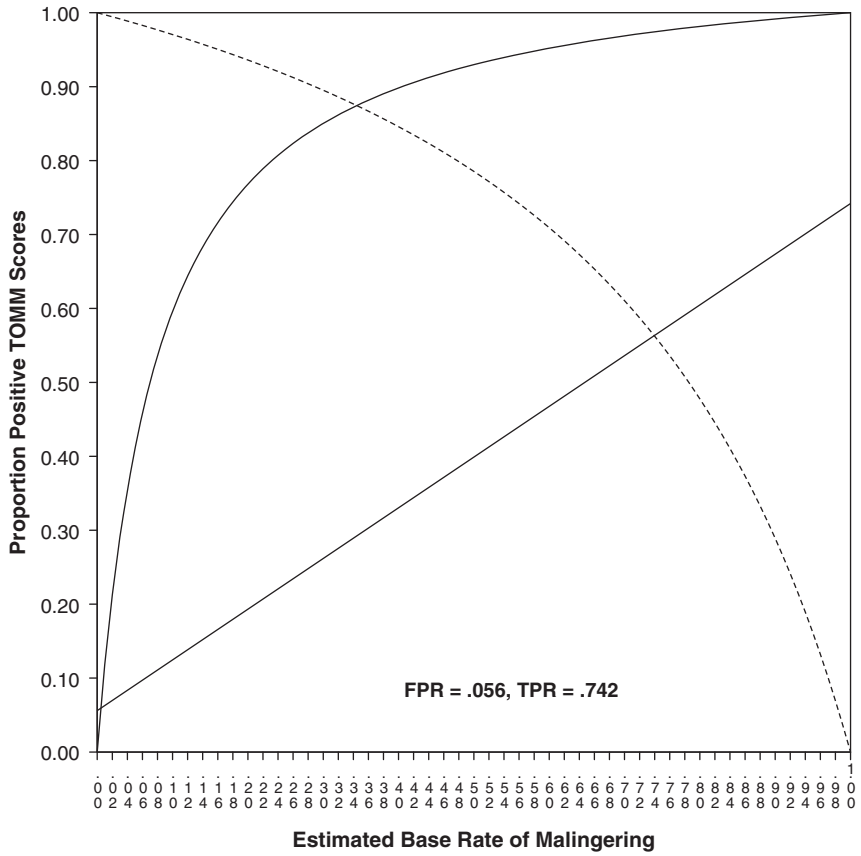
Note: FPR = false positive rate; TPR = true positive rate. Total number of studies included is 24.

through sampling, the standard error of estimation should be routinely reported. A variety of estimation procedures exist. We have used for simplicity of discussion. In some circumstances, other equations will give more accurate error estimates (Agresti & Coull, 1998). Constructing a TVS from multiple studies, however, eliminates the need for separate calculations; the error estimates are an inherent outcome of the regression analysis.

Predictive power is an essential characteristic of decision making, but it is not a parametric value for populations we have considered. Computation of predictive power requires joint consideration of parametric estimates (FPR and TPR) from two distinct *populations* within a local *sample*. Therefore, researchers should not report "the" predictive power of a test. Instead, researchers should report the complete range of predictive power for any specified cut score across the entire BR range. If test researchers report a TVS for each cut score, then test users should be able to identify predictive power of test scores for their local BRs. The TVS makes this a routine and simple exercise. The TVS highlights the curvilinear nature of predictive power and helps test users identify BRs for which the test retains high PPP or NPP. We encourage test users to be familiar with the BR at which the PPP exceeds 50% and to note the NPP that occurs at that BR. A negative test score in the context of low NPP may be just as unhelpful or misleading as a positive test score in the context of a low PPP. This will, of course, depend on the purpose of testing (Griner, Mayewski, Mushlin, & Greenlan, 1981).

Furthermore, the PPL within the TVS can assist clinicians who wish to gain an immediate estimate of

Figure 10
Test Validation Summary for the Test of Memory Malingering (TOMM) Using All Studies Reported in Table 1 but Excluding Studies of Simulators and Challenge Groups



Note: FPR = false positive rate; TPR = true positive rate. As in Figures 7 and 8, in this figure the positive proportion line (diagonal line) is the best-fitting weighted least squares regression of positive test scores on base rates for the studies in Table 2. The solid curved line represents the positive predictive power of the TOMM across all potential base rates. The dashed curved line represents negative predictive power.

their local BR. By taking a sample of test scores from their practice (by random selection or, more easily, by representative consecutive observations), clinicians can readily estimate the local BR for the condition for people who complete those tests. Clinicians need only construct a TVS from published values of FPR and TPR, and once they have computed the *y* value of the PPL (by observation of the rate of positive test scores in their sample), the *x* value (BR for their sample) is readily estimated.

We have already shown how the PPL can be used to estimate more accurately the local BR once a representative rate of positive test results is known. In addition to the description of the TVS, we illustrated the generation of the PPL using WLS regression, which takes into account the sample size from which BRs are estimated. Regarding BR estimation, we have shown, using the regression approach to construction of the PPL, how the *SE*s of the FPR and TPR can be generated. By inference, this should allow investigation into the reliability of BR estimates. Methods for estimation of the reliability of test scores are well known (Nunnally & Bernstein, 1994), although the implications of poorer reliability are often neglected (e.g., Bowden et al., 1988). Of immediate interest is how we can best make reliable estimates of sample BRs in future studies. Greve, Bianchini, Love, et al. (2006) and Greve, Bianchini, and Doane (2006) have developed an important approach in this endeavor. Their use of a stepwise process to consider the probability of

faking and their reporting of cumulative relative frequency distributions at a wide range of cut scores make it possible for other researchers to advance knowledge about testing. Nonetheless, the value of multiple groups with varying probabilities of malingering is diminished by the strictures of known groups validation. Use of mixed group validation with the TVS can attenuate these diminutions.

Good estimates of FPR and TPR and related statistics depend critically on the quality of representative sampling of the respective populations (Brown, 1976). The sensitivity of a test and the associated TPR may appear to increase in a sample in which the more severely unwell individuals are more heavily represented (Hlatky et al., 1984). In the examples used in this article, for example, estimates of sensitivity are likely influenced by the intensity with which individuals intend to malinger. It is interesting, however, that there was not a notable difference in estimates of TPR for the TOMM whether estimating TPR by all samples, including simulators (Figure 8), or using only clinical patients (Figure 9).

Estimates of the sensitivity of a test may also vary with the effect of background variables (e.g., age) if the background variable is associated with clinical characteristics of the abnormal condition (Hlatky et al., 1984). Because small and unrepresentative sampling can have drastic effects of the estimates of population characteristics (Soper, Cicchetti, Satz, Light, & Orsini, 1988), a heavy onus rests on the researcher who wishes to promulgate diagnostic validity statistics for any test to ensure that target populations are carefully sampled and described in detail. A further step to encouraging good estimates of the relevant statistics is to assume that no estimate of classification validity is useful until independently replicated (Mitrushina, Boone, Razani, & D'Elia, 2005). A fundamental tenet of science is independent replication. TPR and FPR estimates represent hypotheses about the construct (or classification) validity of a test in the respective populations and should be subject to the same test of independent replication before being considered to be established.

In the first article on the development of the Validity Indicator Profile, Frederick and Foster (1991) noted the differences in hit rates for "naive" and "informed" simulating malingerers. Had only one sample and not the other been used to estimate population parameter of sensitivity (i.e., for the population of individuals who "have the condition" of malingering), the estimates of TPR would have been notably discrepant. Although we found no striking difference in TOMM estimates when using samples with or without simulators, it is easy to unwittingly influence sampling strategies that lead to inflated or attenuated estimates for the population value of FPR or TPR for malingering tests or any other test. We have stated that TVS analysis assumes that plotted samples are representative of the population. We wish to emphasize that values derived (e.g., PPP and NPP) from a combination of FPR, TPR, and BR are only as good as the estimates of FPR, TPR, and BR. For example, when using the PPL in a TVS to estimate local BR from a sample of tests, it is important to ask whether one's sample is a random and representative sample of the population for which the FPR and TPR are estimated. Unfortunately, many studies of clinical populations do not describe sampling methods in sufficient detail to know whether the sample is representative. It can readily be demonstrated that small, unrepresentative sampling biases estimation of population parameters (Soper et al., 1988). Straus et al. (2005) included a useful checklist for determining the quality of sampling in clinical studies. In analyses such as we did for the TOMM, outliers in a scatter plot of the BR and rate of positive scores may actually identify instances in which a sample has characteristics that argue for exclusion from the analysis.

In closing, we return to our initial discussion of this article. To estimate FPR and TPR, researchers select samples of individuals from Population A and Population B. Distributions of test scores for these samples identify cumulative proportions of individuals with respect to a chosen cut score. The process of estimating FPR and TPR assumes that the samples are pure (probabilities of population membership equal 0 or 1). We believe that this process is much more probabilistic than certain, but standard criterion group methodology does not accommodate probabilities of group membership. Future research should focus on how to better estimate these probabilities. We believe that the TVS methods we have explored can involve mixed groups validation (Frederick, 2000) to overcome this deficiency.

Appendix

SPSS Graphing Program That Can Be Entered Into an SPSS Syntax File and That Will Generate the Test Validation Summary, Given the Entered True Positive Rate (TPR) and False Positive Rate (FPR) Values

```
data list fixed/x 1-3.
begin data.
000
001
Type in separate lines for x-values 002 to 098.
099
100
end data.
COMPUTE TPR = Enter TPR as a decimal value from 0 to 1.
EXECUTE.
COMPUTE FPR = Enter FPR as a decimal value from 0 to 1.
EXECUTE.
COMPUTE N = 100.
EXECUTE.
COMPUTE y = N - x.
EXECUTE.
COMPUTE BR = x/N.
EXECUTE.
COMPUTE NPP = ((1-FPR)*y)/(((1-TPR)*x)+((1-FPR)*y)).
EXECUTE.
COMPUTE PPP = (TPR*x)/((TPR*x)+(FPR*y)).
EXECUTE.
COMPUTE PrPosScore = ((TPR*x)+(FPR*y))/N.
EXECUTE.
GRAPH
/LINE(MULTIPLE)= VALUE(PrPosScore PPP NPP) BY BR.
```

Note: BR = base rate; NPP = negative predictive power; PPP = positive predictive power. Bolded comments are instructions for typing in additional information.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician, 52,* 119-126.

Ashendorf, L., Constantinou, M., & McCaffrey, R. (2004). The effects of depression and anxiety on the TOMM in community-dwelling older adults. *Archives of Clinical Neuropsychology, 19*, 125-130.

Baldessarini, R. J., Finklestein, S., & Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry, 40*, 569-573.

Bolan, B., Foster, J. K., Schmand, B., & Bolan, S. (2002). A comparison of three tests to detect feigned amnesia: The effects of feedback and the measurement of response latency. *Journal of Clinical and Experimental Neuropsychology, 24*, 154-167.

Bowden, S. C., Fowler, K. S., Bell, R. C., Whelan, G., Clifford, C. C., Ritter, A. J., et al. (1998). The reliability and internal validity of the Wisconsin Card Sorting Test. *Neuropsychological Rehabilitation, 8*, 243-254.

Brown, G. W. (1976). Berkson fallacy revisited: Spurious conclusions from patient surveys. *American Journal of Diseases in Children, 130*, 56-60.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Constantinou, M., & McCaffrey, R. J. (2003). Using the TOMM for evaluating children's effort to perform optimally on neuropsychological measures. *Child Neuropsychology, 9*, 81-90.

Cragar, D. E., Berry, D. T. R., Fakhoury, T. A., Cibula, J. E., & Schmitt, F. A. (2006). Performance of patients with epilepsy or psychogenic non-epileptic seizures on four measures of effort. *Clinical Neuropsychologist, 20*, 552-566.

Dawes, R. M. (1967). How clinical probability judgments may be used to validate diagnostic signs. *Journal of Clinical Psychology, 23*, 403-410.

Dawes, R. M., & Meehl, P. E. (1966). Mixed group validation: A method for determining the validity of diagnostic signs without using criterion groups. *Psychological Bulletin, 66*, 63-67.

Delain, S. L., Stafford, K. P., & Ben-Porath, Y. S. (2003). Use of the TOMM in a criminal court forensic assessment setting. *Assessment*, *10*, 370-381.

Donders, J. (2005). Performance on the Test of Memory Malingering in a mixed pediatric sample. *Child Neuropsychology, 11*, 221-227.

Duncan, A. (2005). The impact of cognitive and psychiatric impairment of psychotic disorders on the Test of Memory Malingering (TOMM). *Assessment, 12*, 123-129.

Edens, J. F., Otto, R. K., & Dwyer, T. (1999). Utility of the Structured Inventory of Malingered Symptomatology in identifying persons motivated to malinger psychopathology. *Journal of the American Academy of Psychiatry and Law, 27*, 387-396.

Elwood, R. W. (1993). Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review, 13*, 409-419.

Etherton, J. L., Bianchini, K. J., Heinly, M. T., & Greve, K. W. (2006). Pain, malingering, and performance on the WAIS-III Processing Speed Index. *Journal of Clinical and Experimental Neuropsychology, 28*, 1218-1237.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.

Frederick, R. I. (2000). Mixed group validation: A method to address the limitations of criterion group validation in research on malingering detection. *Behavioral Sciences and the Law, 18*, 693-718.

Frederick, R. I., & Foster, H. G. (1991). Multiple measures of malingering on a forced-choice test of cognitive ability. *Assessment, 3*, 596-602.

Gavett, B. E., O'Bryant, S. E., Fisher, J. M., & McCaffrey, R. J. (2005). Hit rates of adequate performance based on the Test of Memory Malingering (TOMM) Trial 1. *Applied Neuropsychology, 12*, 1-4.

Gervais, R. O., Rohling, M. L., Green, P., & Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Archives of Clinical Neuropsychology, 19*, 475-487.

Gierok, S. D., Dickson, A. L., & Cole, J. A. (2005). Performance of forensic and non-forensic adult psychiatric inpatients on the Test of Memory Malingering. *Archive of Clinical Neuropsychology, 20*, 755-760.

Green, P. (2007). Spoiled for choice: Making comparisons between forced choice effort tests. In K. B. Boone (Ed.), *Detection of noncredible cognitive performance* (pp. 50-77). New York: Guilford.

Greve, K. W., Bianchini, K. J., & Doane, B. M. (2006). Classification accuracy of the Test of Memory Malingering in traumatic brain injury: Results of a known-groups analysis. J*ournal of Clinical and Experimental Neuropsychology, 28*, 1176-1190.

Greve, K. W., Bianchini, K. J., Love, J. M., Brennan, A., & Heinly, M. T. (2006). Sensitivity and specificity of MMPI-2 validity scales and indicators to malingered neurocognitive dysfunction in traumatic brain injury. *Clinical Neuropsychologist, 20*, 491-512.

Griner, P. F., Mayewski, R. J., Mushlin, A. I., & Greenlan, P. (1981). Selection and interpretation of diagnostic tests and procedures. *Annals of Internal Medicine, 94*, 557-592.

Haber, A. H., & Fichtenberg, N. L. (2006). Replication of the Test of Memory Malingering (TOMM) in a traumatic brain injury and head trauma sample. *Clinical Neuropsychologist, 20*, 524-532.

Hill, S. K., Ryan, L. M., Kennedy, C. H., & Malamut, B. L. (2003). The relationship between measures of declarative memory and the Test of Memory Malingering in patients with and without temporal lobe dysfunction. *Journal of Forensic Neuropsychology, 3*, 1-18.

Hlatky, M. A., Pryor, D. B., Harrell, F. E. J., Califf, R. M., Mark, D. B., & Rosati, R. A. (1984). Factors affecting sensitivity and specificity of exercise electrocardiography. *American Journal of Medicine, 77,* 64-71.

Lewis, J. L., Simcox, A. M., & Berry, D. T. R. (2002). Screening for feigned psychiatric symptoms in a forensic sample by using the MMPI-2 and the Structured Inventory of Malingered Symptomatology. *Psychological Assessment, 14*, 170-176.

McCaffrey, R. J., O'Bryant, S. E., & Fisher, J. M. (2003). Correlations among the TOMM, Rey-15, and MMPI-2 validity scales in a sample of TBI litigants. *Journal of Forensic Neuropsychology, 3*, 45-53.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194-216.

Merckelbach, H., & Smith, G. P. (2003). Diagnostic accuracy of the Structured Inventory of Malingered Symptomatology (SIMS) in detecting instructed malingering. *Archives of Clinical Neuropsychology, 18*, 145-152.

Miller, H. A. (2001). *The Miller Forensic Assessment of Symptoms Test: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*, 1094-1102.

Moore, B. A., & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury, 18*, 975-984.

Moss, A., Jones, C., Fokias, D., & Quinn, D. (2003). The mediating effects of effort upon the relationship between head injury severity and cognitive functioning. *Brain Injury, 17*, 377-387.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Bryant, S. E., Hilsabeck, R. C., Fisher, J. M., & McCaffrey, R. J. (2003). Utility of the Trail Making Test in the assessment of malingering in a sample of mild traumatic brain injury litigants. *Clinical Neuropsychologist, 17*, 69-74.

O'Bryant, S. E., & Lucas, J. A. (2006). Estimating the predictive value of the Test of Memory Malingering: An illustrative example for clinicians. *Clinical Neuropsychologist, 20*, 533-540.

Powell, M. R., Gfeller, J. D., Hendricks, B. L., & Sharland, M. (2004). Detecting symptom- and test-coached simulators with the Test of Memory Malingering. *Archives of Clinical Neuropsychology, 19*, 693-702.

Rees, L. M., Tombaugh, T. N., & Boulay, L. (2001). Depression and the Test of Memory Malingering. *Archives of Clinical Neuropsychology, 16*, 501-506.

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*, 10-20.

Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *Structured Interview of Reported Symptoms: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Rogers, R., Jackson, R. L., Sewell, K. W., & Harrison, K. S. (2004). An examination of the ECST-R as a screen for feigned incompetency to stand trial. *Psychological Assessment, 16*, 139-145.

Rogers, R., Tillbrook, C. E., & Sewell, K. W. (2004). *Evaluation of Competency to Stand Trial–Revised.* Odessa, FL: Psychological Assessment Resources.

Rohling, M. L., & Boone, K. B. (2007). Future directions in effort assessment. In K. B. Boone (Ed.), *Detection of noncredible cognitive performance* (pp. 453-470). New York: Guilford.

Rorer, L. G., & Dawes, R. M. (1982). A baserate bootstrap. *Journal of Consulting and Clinical Psychology*, *50*, 419-425.

Rosenfeld, B., Sands, S. A., & van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology, 15*, 349-359.

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *Clinical Neuropsychologist, 13,* 545-561.

Soper, H. V., Cicchetti, D. V., Satz, P., Light, R., & Orsini, D. L. (1988). Null hypothesis disrespect in neuropsychology: Dangers of alpha and beta errors. *Journal of Clinical and Experimental Neuropsychology, 10*, 255-270.

Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3d ed.). Edinburgh, UK: Elsevier Churchill-Livingstone.

Tan, J. E., Slick, D. J., Strauss, E., & Hultsch, D. F. (2002). How'd they do it? Malingering strategies on symptom validity tests. *Clinical Neuropsychologist, 16*, 495-505.

Teichner, G., & Wagner, M. T. (2004). The Test of Memory Malingering (TOMM): Normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Archives of Clinical Neuropsychology, 19*, 455-464.

Tombaugh, T. N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment, 9,* 260-268.

Vickery, C. D., Berry, D. T. R., Dearth, C. S., Vagnini, V. L., Baser, R. E., Crager, D. E., et al. (2004). Head injury and the ability to feign neuropsychological deficits. *Archives of Clinical Neuropsychology, 19,* 37-48.

Weinborn, M., Orr, T., Woods, S. P., Conover, E., & Feix, J. (2003). A validation of the Test of Memory Malingering in a forensic psychiatric setting. *Journal of Clinical and Experimental Neuropsychology, 25*, 979-990.

Widows, M. R., & Smith, G. P. (2005). *Structured Inventory of Malingered Symptomatology: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate, Washington, DC. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Woods, S. P., Weinborn, M., & Lovejoy, D. W. (2003). Are classification accuracy statistics underused in neuropsychological research? *Journal of Clinical and Experimental Neuropsychology, 25*, 431-439.

Zakzanis, K. K. (1998). Brain is related to behavior ($p < .05$). *Journal of Clinical and Experimental Neuropsychology, 20*, 419-427.

Zakzanis, K. K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: Formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology, 16*, 653-667.