*J Head Trauma Rehabil*
Vol. 24, No. 2, pp. 105–122


# Evaluating Constructs Represented by Symptom Validity Tests in Forensic Neuropsychological Assessment of Traumatic Brain Injury

*Richard I. Frederick, PhD; Stephen C. Bowden, PhD*

This study uses a new method to summarize diagnostic validity information to explore which constructs are captured by malingering tests. The Test Validation Summary applies mixed-groups validation to investigate the meaning of test constructs and to estimate test classification characteristics when test validation groups are not "pure" criterion groups (ie, "compliant" vs "malingering"), but members have variable probability of malingering. The method permits the use of tests with relatively low validity to validate tests of greater validity. In our initial analysis, we argue that the Rey 15-Item Memory Test is best construed as an "intention test" (capturing the intention of test takers when taking a test) as opposed to an "effort test." Using the Test Validation Summary and mixed-groups validation, we demonstrate that as an indicator of "intention to feign cognitive impairment," the Rey 15-Item Memory Test has estimated false-positive rate (FPR) = 0.02 and true-positive rate (TPR) = 0.57. We then explore the meaning of failure on the Word Memory Test (WMT), which uses a dichotomous classification of performance as valid or invalid. Although the WMT is commonly referred to as an "effort test," we argue that it likely captures both "intention" and "effort" but collapses this information into a single dichotomous classification of symptom validity. We demonstrate that, as a result of this dichotomous classification process, the WMT likely has a problematic FPR. In our analysis of previously published WMT data, the WMT FPR is estimated at 0.12 when there is no predisposition to perform poorly but rises dramatically and unrealistically as the predisposition to perform poorly increases. We compare these findings to those of the Validity Indicator Profile (VIP), which captures both intent and effort to classify 4 different sorts of response styles in cognitive testing. In our analyses, the VIP demonstrates that FPR = 0 and TPR = 0.86 when the construct being measured is "intent to perform poorly," and reveals that FPR = 0.06 and TPR = 0.63 when the construct being measured is "inconsistent responding" or "poor effort." We were able to demonstrate for the VIP the same "oversensitivity" shown by the WMT when the VIP was interpreted only as a dichotomous classification test. These results indicate that researchers who attempt to generate classification characteristics for malingering tests must carefully consider what constructs are being captured by the test. **Keywords:** *forensic neuropsychology*, *feigned cognitive impairment*, *malingering assessment*, *test validation*

CLINICIANS conducting psychological evaluations to address legal or compensatory issues for patients with traumatic brain injuries (TBIs) are obliged to examine symptom validity.[1] One consequence of the increasing recognition of potential manipulation of cognitive and psychosocial symptoms by examinees is the proliferation of symptom validity tests. Forensic neuropsychological assessments that do not fully address issues of symptom validity are considered to be incomplete.[2] However, the diagnostic quality of symptom validity tests varies widely and presents a significant challenge for accurate assessments (Table 1).[3,4]

This study explores issues related to test classification accuracy of some symptom validity tests. We introduced the Test Validation Summary[5] (TVS) as a method for graphing and estimating the false-positive rate (FPR) and true-positive rate (TPR) of a test cut score. The TVS plots rates of positive test scores as a function of the base rate (BR) of a condition in a mixed sample taken from 2 populations: Population A comprises all individuals with a condition, while Population B comprises all individuals without the condition. When a sample is "pure," having only Population A members, the rate of positive test scores estimates the TPR. When a sample has only Population B members, the rate of positive test scores estimates the FPR. When a sample is mixed, the TVS

***Author Affiliations:*** *Department of Psychology, US Medical Center for Federal Prisoners, Springfield, Missouri (Dr Frederick); and Department of Psychology, University of Melbourne, Parkville, Victoria, Australia (Dr Bowden).*

***Corresponding Author:*** *Richard Frederick, PhD, Department of Psychology, US Medical Center for Federal Prisoners, 1900 W Sunshine St, Springfield, MO 65807 (rickfrederick@gmail.com).*

**TABLE 1** *Acronyms*

| | |
|---|---|
| BR | Base rate—the proportion of individuals in a sample who have the condition |
| CGV | Criterion groups validation—a special case of MGV, in which the varying base rates of the condition in samples is restricted to BR = 0 and BR = 1. |
| FPR | False-positive rate—the probability that a person who does not have a condition will generate a test score which is consistent with having the condition |
| MGV | Mixed groups validation—using "mixed" groups (groups with varying base rates of the condition) to validate a chosen test cut score |
| MMPI-2 | Minnesota Multiphasic Personality Inventory-2 |
| NPP | Negative positive power—the probability that a person who has a test score consistent with not having a condition actually does not have the condition |
| PPP | Positive predictive power—the probability that a person who has a test score consistent with having a condition actually has the condition |
| PPL | Positive proportion line—a straight-line display in the TVS that reflects the increasing proportion of positive test scores as the BR of the condition increases in a sample. The end points of the PPL are the FPR and the TPR of the chosen test cut score. |
| RMT | Rey 15-Item Memory Test |
| TBI | Traumatic brain injury |
| TOMM | Test of Memory Malingering |
| TPR | True-positive rate—the probability that a person who has a condition will generate a test score which is consistent with having the condition |
| TVS | Test Validation Summary—a graphical display of the relationship among FPR, TPR, PPP, NPP, and BR |
| VIP | Validity Indicator Profile |
| VRIN | Variable Response Inconsistency |
| WLS | Weighted least squares |
| WMT | Word Memory Test |

estimates the proportion of Population A and B members from the rate of positive test scores.

As an example, Figure 1 displays a diagonal line, which is the foundation of the TVS, the positive proportion line (PPL). The *x* axis represents the range of the BR, that is, the proportion of people in any sample with the condition. The *y* axis represents the proportion of test scores that are positive. Points on the positive proportion line (PPL) indicate the proportion of people with positive test scores at a given BR. For example, at the left end point of the PPL, the BR of population A is 0. The rate of positive test scores at BR = 0 is 0.16, which is the rate of positive test scores for individuals without the condition, the FPR. At the right end point of the PPL, the BR = 1, all members of the sample have the condition–the rate of positive test scores estimates the TPR = 0.50. In Figure 2, a mixed sample generates 33% positive scores. The rate of population A members is estimated at 0.50. This makes sense: Rate of positive scores = (FPR)(1 – BR) + (TPR)(BR) = (0.16)(1 – 0.50) + (0.50)(0.50) = 0.08 + 0.25 = 0.33.

Test developers are interested in FPR and TPR. Test users (eg, clinicians) should be interested in predictive power. Positive predictive power (PPP) is the probability that a person has the condition of interest when generating a positive score on the test. Negative predictive power (NPP) is the probability that a person does not have the condition of interest given a negative score on the test. Figure 3 represents a completed TVS. A fundamental contribution of the TVS is that, when given the FPR and TPR of a test, the TVS shows the range of PPP (solid curved line) and NPP (dashed line) across the range of BRs in which the test might be administered. Given the curvilinear nature of PPP and NPP, it is often difficult for clinicians to informally estimate the PPP and NPP, and the computations are slightly cumbersome. The TVS obviates the need for hand computations. As shown in Figure 4, once the TVS is drawn, knowing the percentage of positive scores in a sample leads to estimation of the sample BR, PPP, and NPP. Conversely, a good estimate of BR allows estimation of PPP, NPP, and proportion positive scores.

In our previous article,[5] we demonstrated that a TVS could be derived from weighted least squares (WLS) regression analysis of a number of studies that reported TPR and FPR for a given cut score. (The number of participants in each study is the weighted element of regression.) Such estimates often reflect a wide range of values. By collapsing research samples that had been intended to be pure ("known" groups analysis) into mixed samples or by estimating BRs for some clinical samples, we were able to generate a wide range of BRs and rates of positive test scores for the Test of Memory Malingering[6] (TOMM), a commonly used symptom validity measure in forensic neuropsychology.[7,8] From 25 research reports about the TOMM, we plotted 37 pairs of (BR, +) and,

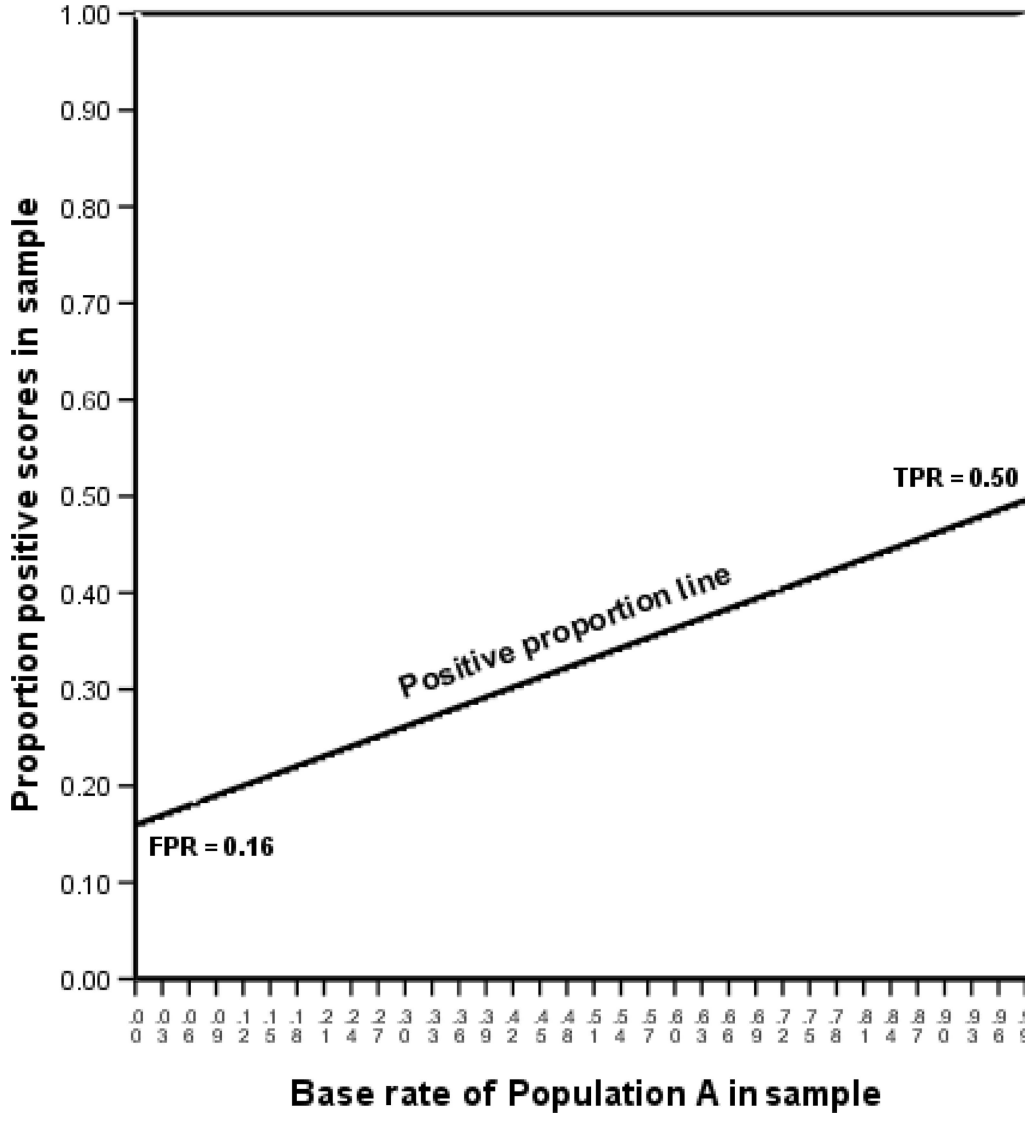


**Figure 1.** Initial plotting of the Test Validation Summary (TVS). The straight diagonal line is the positive proportion line (PPL), which demonstrates the range of the proportion of positive test scores for a hypothetical test used to classify test takers as members of Population A (who have the condition of interest) or Population B (who do not). Values plotted on the $x$ axis reflect the rate of Population A members in the tested sample, the local base rate. 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), $y = 0.16$ (ie, FPR = 0.16). When $x = 1$ (when there are no members of Population B, BR = 1), $y = 0.50$ (ie, TPR = 0.50). Connecting these 2 points, (0, 0.16) and (1, 0.50), plots the PPL.

by WLS regression, derived the end points of the PPL (the PPL is the least-squares regression line). Estimated FPR = 0.052 (SE = 0.021), and estimated TPR = 0.777 (SE = 0.061). When we removed the simulation studies from the analysis, FPR was estimated at 0.056 (SE = 0.025), and TPR was estimated at 0.742 (SE = 0.093). This latter set of results is shown in Figure 5. Note also that the analysis yields standard errors (SEs) for FPR and TPR, an interval estimation metric typically neglected in the reporting of FPR and TPR,[9] although interval estimation has had official imprimatur in psychology for almost a decade.[10] Only 2 clinical studies in our sample had BR > 0.50. As a consequence, many fewer points contributed to the estimate of TPR versus FPR, hence the larger SE.

Although the TOMM has sometimes been referred to as an "effort test,"[11] we note that the task demand is so minimal that even significantly impaired,[6] very young,[12] or very old individuals[13] have little difficulty in meeting the cutoff score for compliant performance. We choose, then, to construe the TOMM as an "intention" test, one that requires only the intent to perform well for

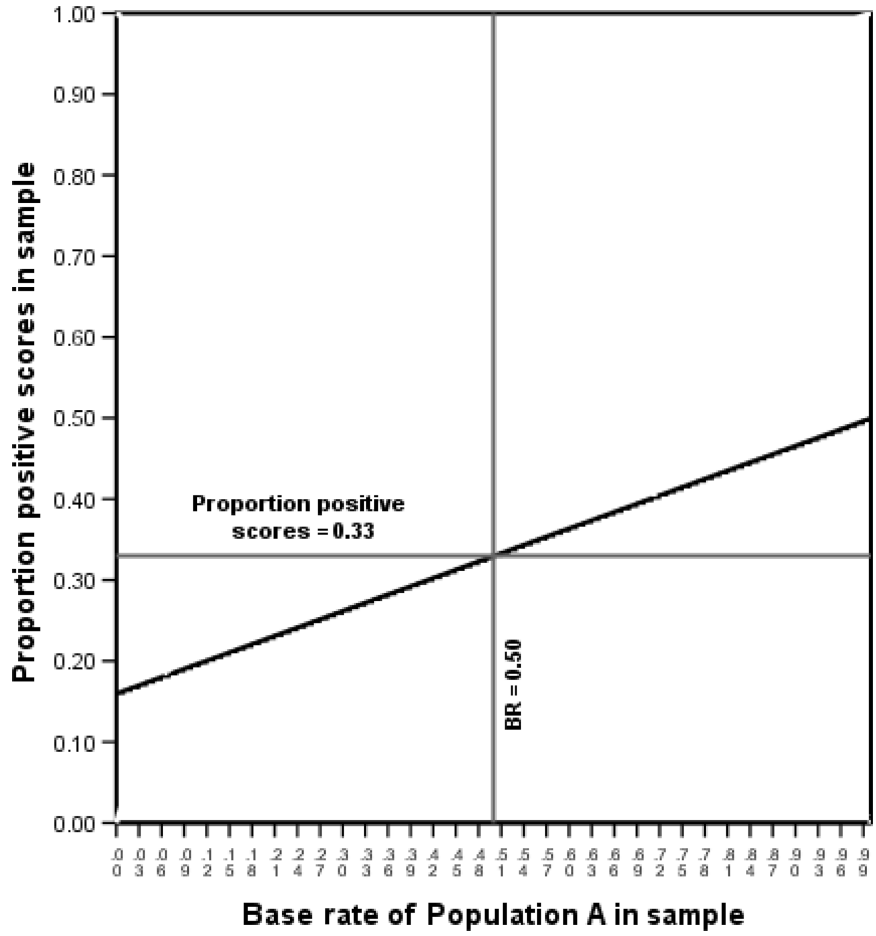


**Figure 2.** A sample of test takers in which 33% of test scores are positive. By plotting this value (0.33) on the *y* axis, one can locate the corresponding *x* value on the PPL. Here, the *x* value = 0.50. This value represents the best estimate of the BR of Population A members in the sample, given the test's reported values of FPR and TPR.

successful performance. That is, the task demand for successful performance on the TOMM is likely so minimal as to preclude the measurement of effort. In this construction, the Figure 5 estimates of TOMM FPR (0.056) and TPR FPR (0.742) mean that when everyone (including severely impaired individuals) in a sample intends to respond correctly to all items (ie, BR = 0), 5.6% of them fail the TOMM, and when no one in a sample intends to respond correctly to all items (ie, BR = 1), then 74.2% of them fail the TOMM.

In this article, we wish to explore the utility of the TVS in construct validation of tests commonly encountered in forensic neuropsychological examinations. In our first analysis, we use WLS regression to evaluate the meaning of the construct underlying the Rey 15-Item Memory Test (RMT).[14–16] The purpose of the analysis is 2-fold: (1) we wish to show how the TVS method can clarify the meaning of constructs and (2) we wish to use the results in a later analysis in the article in which we show that a less valid test can be used to validate a more valid test. In our second analysis, we explore whether 1 or 2 constructs underlie the Word Memory Test (WMT).[17] In our third analysis, we demonstrate that 2 constructs underlie the Validity Indicator Profile (VIP).[18]

## REY 15-ITEM MEMORY TEST

To consider the possibilities of clarifying the constructs captured by a test, we first consider the RMT (see Frederick[15] for a translation of Rey[14]). The RMT requires test takers to memorize 15 printed items over a short period of visual exposure and then to reproduce the items on a blank sheet of paper. The score is the number of correctly reproduced items. The RMT is an easy test for cooperative individuals, even if they have significant impairment, but, nevertheless, in many published instructions, the test is presented as "a very difficult test of memory." Many cagey examinees may instantly recognize such a statement as disingenuous and suspect a trap, which may lead them to take the test honestly. Consequently, some researchers[15,19] have recommended that examiners refrain from this instructional set as a

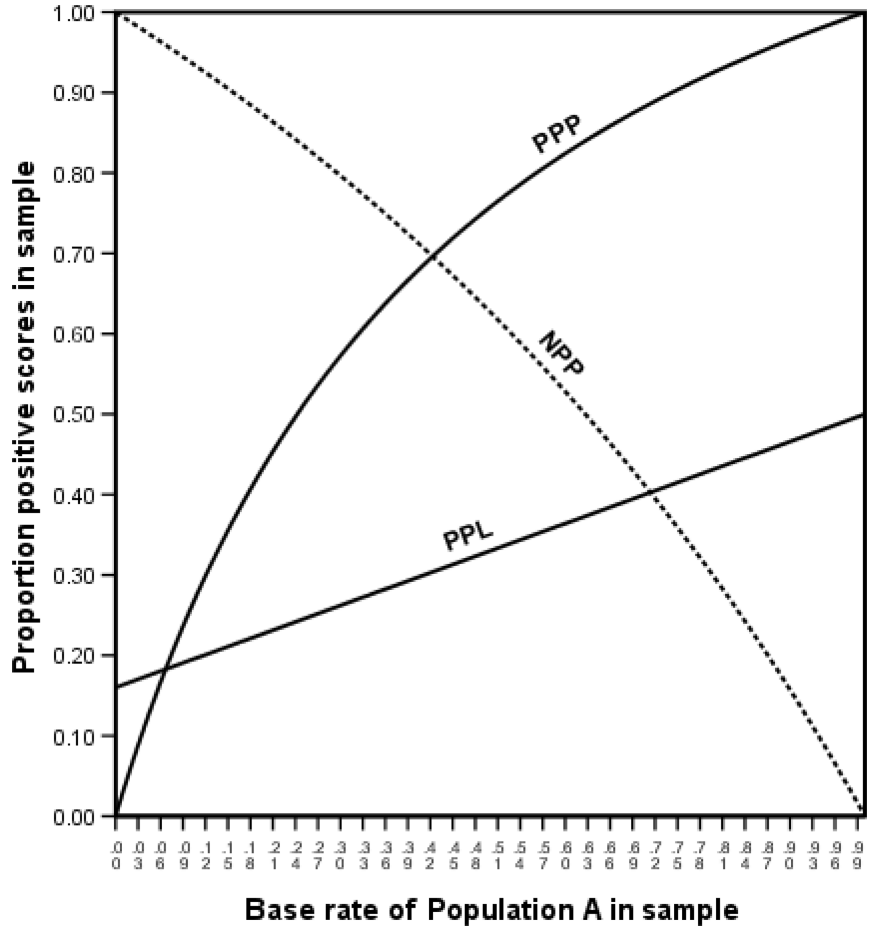


**Figure 3.** A complete TVS for the hypothetical test. Now are included plots for positive predictive power (PPP, solid curve) and negative predictive power (NPP, dotted curve).

means to potentially increase the effectiveness of the test.

Frederick[20] reported the rates of RMT positive test scores generated by 723 criminal defendants undergoing mental health evaluations by order of a federal court. Although there was a low rate of diagnosed TBI among the defendants, cognitive capacity is a fundamental component of competency to stand trial and criminal responsibility and validity issues must be addressed. Cochrane et al[21] reported that the rate of malingering among 1710 criminal defendants undergoing examination was 10.9%. Defendants in Frederick[20] were assessed with a number of symptom validity tests, including the RMT. Prior to any testing, each criminal defendant was rated by the primary clinician with the query "What is the probability that the defendant will feign cognitive impairment?" Ratings were assigned on a scale of 0 to 100. Although not reported in Frederick,[20] an additional rating was obtained from primary clinicians. Before any tests were administered, clinicians, when making ratings of the probability that a defendant would feign cognitive impairment, were simultaneously asked to rate the probability that a defendant would feign memory impairment using the same scale, 0 to 100. These data are provided now by one of the authors of this article (R.I.F.).

To generate multiple points for regression with a reasonable number of defendants in each group, we divided the 723 defendants into 28 groups of 25 and 1 group of 31 after we sorted them by ratings. For each of the 29 groups, we estimated 2 sets of BR by the mean probability rating assigned by clinicians for (1) feigned cognitive impairment and (2) feigned memory impairment. We then determined the proportion of scores in each group for which RMT < 9, the most common cut score for RMT.[15,20] The scatterplots of these values ($x$ = BR, $y$ = proportion RMT < 9; ■ = cognitive faking, ○ = memory faking) are shown in Figure 6. Two PPLs were derived from WLS regression analysis. In Figure 6, we can compare PPLs for RMT < 9 on the basis of the predictor variable. Using the probability of feigned memory impairment as the predictor, RMT < 9 is estimated to have FPR = 0.040 (SE = 0.013; 95% CI = 0.015–0.065) and TPR = 0.467 (SE = 0.040, 95% CI = 0.389–0.545). Using the probability of feigned cognitive impairment

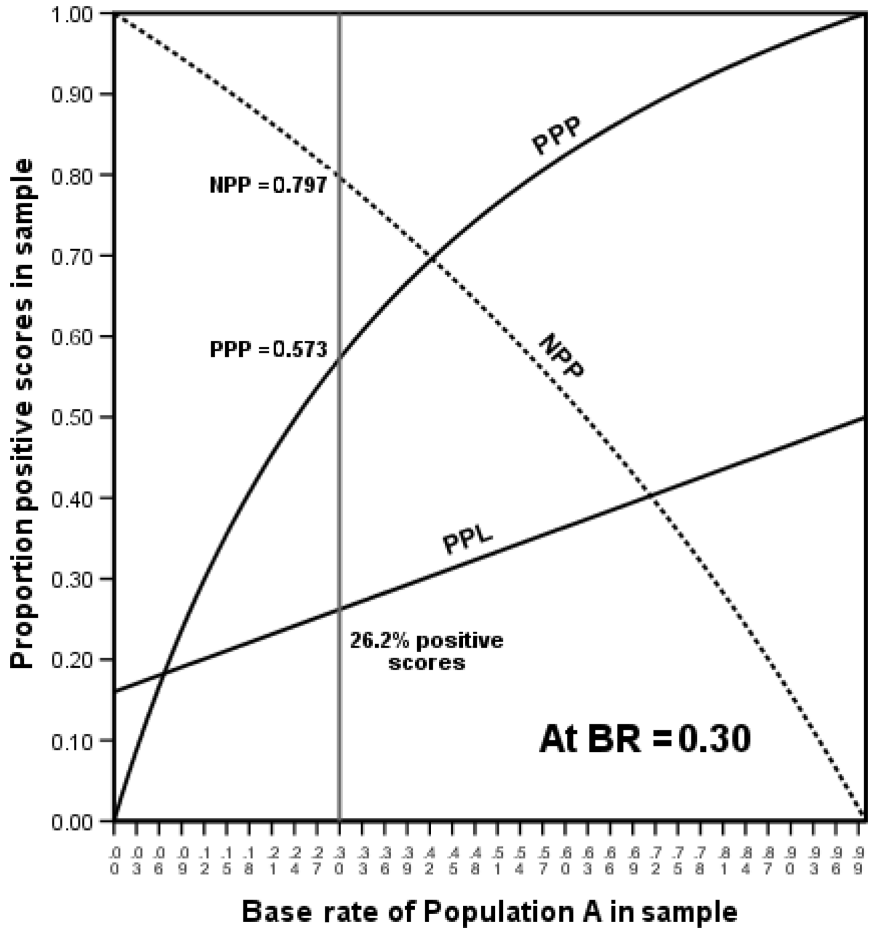


**Figure 4.** The figure shows how to estimate PPP, NPP, and the proportion of positive tests likely to be seen when the sample base rate = 0.30. Conversely, as in Figure 2, if a random sample of persons taking this test generates 26.2% positive scores, then the BR is estimated at 0.30, and the PPP and NPP estimates for this sample are easily derived by observation.

as the predictor, RMT < 9 is estimated to have FPR = 0.025 (SE = 0.002; 95% CI = 0.021–0.029) and TPR = 0.574 (SE = 0.044; 95% CI = 0.488–0.660). The value for TPR for each method is not within the 95% CI for TPR generated by the other method.

In this regression of mean RMT failures (positive RMT scores) on mean probability judgments by clinicians (likelihood of feigning cognitive impairment), we note that, across our 28 groups, variation in mean probability judgments accounted for 80.6% of the variation in mean RMT failures (adjusted $R^2 = 0.806$, multiple $R = 0.902$); $F_{1,27} = 117.4$, $P < .001$. This analysis speaks to the construct assessed by the RMT. The only difference in how the PPLs in Figure 6 were generated was the change in the word *memory* versus *cognitive* in the rating instruction. One hypothesis from a comparison of the 2 PPLs in Figure 6 is that it means something different for a test taker to feign cognitive impairment as opposed to memory impairment, particularly on the RMT. In other words, the face valid characteristics of the RMT may be more sensitive to individuals who wish to feign cognitive impairment as opposed to memory impairment, and we suggest that this is a form of construct validation for the RMT made possible by TVS analysis. We believe that this finding has implications for the validation of other malingering tests. In particular, we are interested in exploring the routine use of the rubric "effort test" to refer to symptom validity tests. It is our belief that "intention" (what one intends to do in testing) and "effort" (how hard one works to fulfill one's intention) are 2 separate constructs to be accounted for by tests. Because tests like the RMT have essentially no task demand, they cannot reasonably be considered "effort" tests,[22] and we believe they should be labeled as "intention" tests. We continue our exploration of this idea with the WMT.

## WORD MEMORY TEST

The WMT is described as a test of "cognitive effort."[17] The WMT presents a list of words for memorization. Success at learning this list is examined by a 2-alternative, forced-choice recognition at immediate and delayed

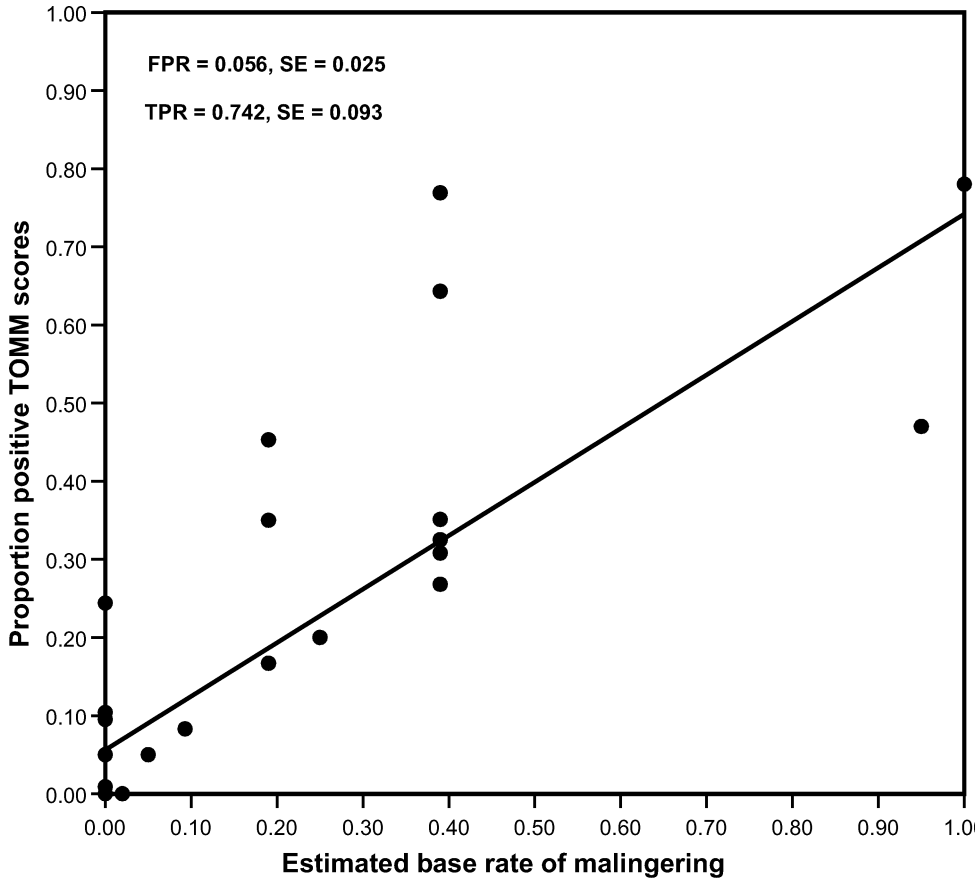


**Figure 5.** A TVS constructed for the Test of Memory Malingering (TOMM) constructed by WLS regression of 24 studies (no simulation studies) in which the base rate of malingering in the study is the independent variable and the rate of positive TOMM scores is the dependent variable. Based on the end points of the positive proportion line, the TOMM FPR is estimated at 0.056 (SE = 0.025) and the TPR is estimated at 0.742 (SE = 0.093). See Frederick and Bowden[5] for more discussion and explanation of this figure.

intervals, as well as delayed free recall. The failure criterion is a score of less than 35 of 40 on any one of the immediate or delayed recall trials, or on a recall consistency score. The test authors have made strong claims regarding the value of the WMT as a test of poor effort, describing the test as "unique among symptom validity tests because of its extensive validation in clinical forensic settings, rather than relying on simulation research with healthy volunteers."[23(p97)] Green has described "effort testing" as a means of identifying malingering: "... failure on effort testing indicates insufficient effort to produce valid test results. In such cases, symptom exaggeration is invariably present."[24(p235)]

Flaro et al[25] claimed that failure rate on the WMT is "23 times higher in [patients with] mild brain injury" than in parents seeking custody, with the implication of neuropsychological performance of patients with mild TBI as not being valid (1.7% failure rate for parents seeking custody vs 40% failure rate for persons seeking compensation for mild TBI). According to Green, failure on the WMT is properly interpreted to mean that "... there is a very high probability that other test scores from the same person significantly underestimate the person's actual abilities, owing to poor effort. This is especially true for memory tests but also for most other tests .... *There are very few false positives*" (emphasis added).[26(p8)] Green also stated:

> If a person fails the WMT effort measures, they fall into one of two categories. In most cases, ... the person probably did not make sufficient effort to produce valid results. It is likely that the person's test results will overestimate any impairment actually present. Alternatively, as with other effort measures, in relatively few cases, it may be concluded that the person was actually unable to pass the WMT effort measures because of very severe and widespread cognitive impairment.[26(p40)]

In addition, the author also stated as follows:

> It is important to point out that failure on the WMT effort measures does not mean that a person will automatically be labeled as a malingerer. This is such a pejorative and emotionally

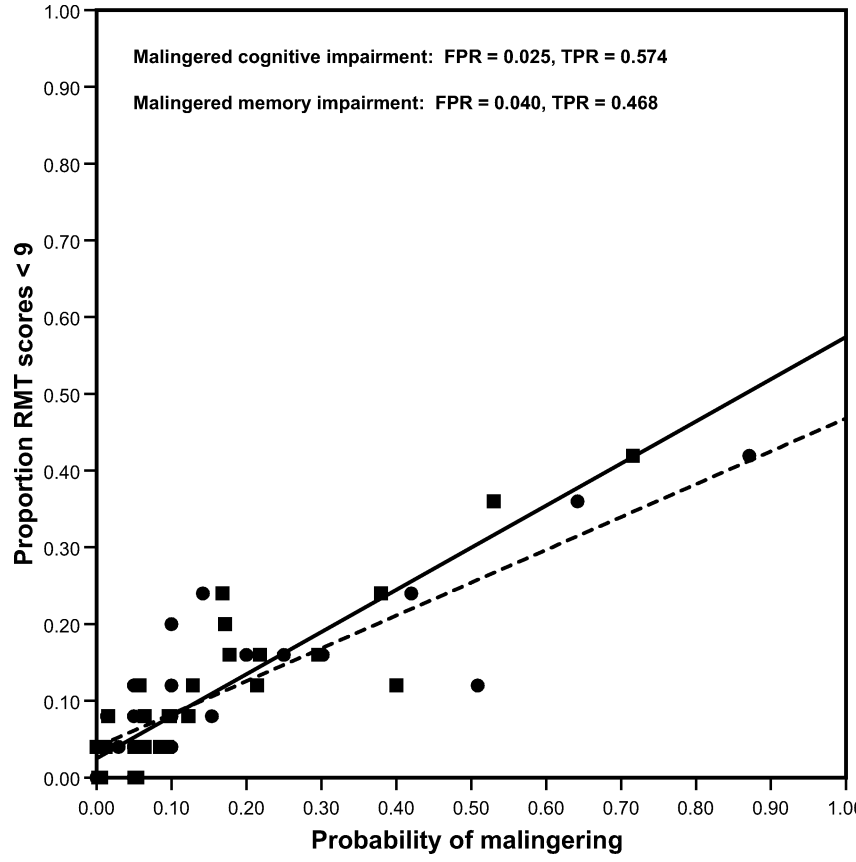


**Figure 6.** A comparison of PPLs for the Rey 15-Item Memory Test (RMT), when the chosen cut score is fewer than 9 items reproduced. Two PPLs are shown, derived from WLS regression. The predictor variable in these PPLs is a clinical judgment about malingering (ranging from 0 to 100; here expressed as a proportion) for 731 criminal defendants. The differences in the PPLs is attributed to whether the clinical rating concerned the probability of faked cognitive impairment (squares and solid line) or faked memory impairment (circles and dashed line). Based on the type of clinical rating, the RMT cut score of 9 generates statistically significantly different TPR values. Based on the rating of predisposition of criminal defendants to feign cognitive impairment, the estimated TPR value is 0.574 (SE = 0.044). Based on the ratings of predisposition of criminal defendants to feign memory impairment, the estimated TPR is 0.468 (SE = 0.040).

laden term that . . . [i]t may be most prudent to report WMT findings operationally and to keep any discussions of motivation separate from WMT results and the validity of other results.[26(p40)]

Nevertheless, the "pass-fail" structure of the WMT and the dichotomization of "failures" as "insufficient effort" or "severe impairment" belie this attempt to prevent description of the WMT as a "malingering test." The WMT is routinely considered to be a malingering test, and failures are considered to represent instances of malingering, even a gold standard of malingering.[27] Indeed, the title of a 1999 article by Green et al[28] is "Detecting Malingering in Head Injury Litigation With the Word Memory Test."

It seems to us that Green's[11,26] use of the term *poor effort* effectively incorporates a variety of response styles, which, on cognitive tests, can include an intent to perform poorly, a lack of intention to perform well, a compromised intent to perform well (eg, half-heartedness), inattentiveness, distractibility, and carelessness. The ambiguous nature of Green's interpretation for the dichotomous classification scheme of the WMT makes no distinction among these potential response styles. As we have shown above, a common interpretation of a failure on the WMT is simply "malingering."

In contrast, Frederick and colleagues[18,29–31] proposed a model of test response validity in which performance is best represented by a cross-classification of intent ("intends to respond correctly" vs "does not intend to respond correctly") with effort (intensity of application toward a goal on a continuum from low to high). Frederick[18,29] uses the term *effort*, markedly differently from Green.[26] As shown in Figure 7, Frederick's[18,29] cross-classifications result in 4 characterizations of test performance: (1) high effort to respond correctly–compliant (valid performance), (2) low effort to

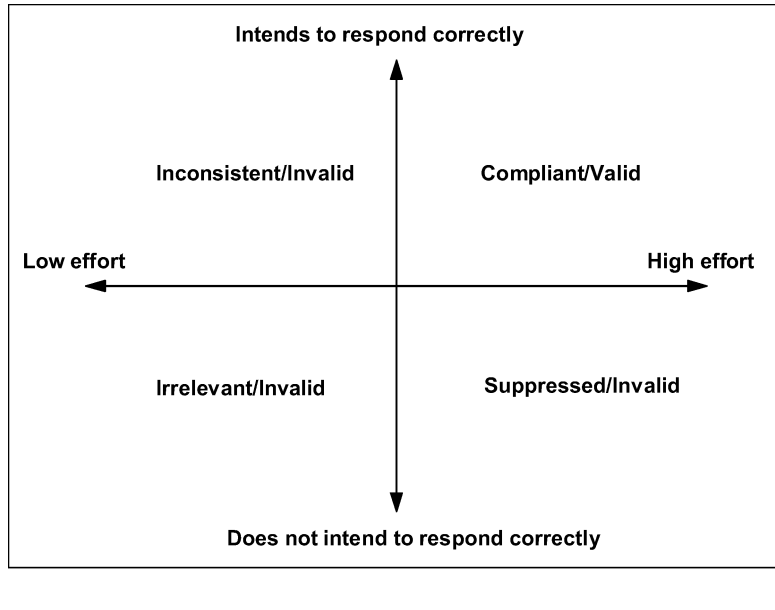


**Figure 7.** Potential categories of response style on cognitive test based on a cross-classification of "intent" and "effort." *Intent* refers to the goal of the test taker in completing test items (either intends to respond correctly or does not intend to respond correctly). Effort refers to the intensity of application of ability to solve problems (a continuum from low to high). The cross-classification yields one valid response style, Compliant, which reflects high effort from an individual who intends to respond correctly. All other cross-classifications (Inconsistent, Irrelevant, and Suppression) reflect invalid response styles; test performance cannot be considered to accurately represent the best abilities of the test taker.

respond correctly–inconsistent (invalid performance), (3) low effort with no intent to respond correctly–irrelevant (invalid performance), and (4) high effort with no intent to respond correctly–suppression (invalid performance). *Suppression* means that the test taker intentionally suppresses true ability and supplies incorrect responses.[18]

Whichever response style the WMT is capturing, it nonetheless delivers a dichotomous decision. Green[32] has compared the results of the WMT with another dichotomous-decision test, the TOMM. As we noted above, we also consider the TOMM to be an "intention test." So, although Green[32] ostensibly compares the 2 tests as "effort tests," we think that a more precise description is that Green[32] is comparing the WMT (whichever constructs it is actually capturing) to an "intention" test.

Green[32] reviewed Gervais et al,[33] who reported results of the WMT and the TOMM for 519 consecutively referred nonhead injury disability claimants, who were seeking worker's compensation for long-term personal injury disability. Participants were assigned to a "Psychological group" ($n = 326$) if they were referred to determine suitability for long-term disability benefits or "to evaluate the extent of pain or other psychological damages in the context of personal injury litigation."(p476) Participants were assigned to a "Vocational group" ($n = 193$) if they had been referred to "determine their suitability for vocational retraining because they could not return to their former occupation."(p476)

In Table 2 of their study, Gervais et al[33](p479) reported the following rates of failure for the WMT and TOMM: psychological group, WMT = 0.43, TOMM = 0.17; vocational group, WMT = 0.12, TOMM = 0.01. The authors concluded that these differential rates of failure between the TOMM and the WMT mean that the "tests vary substantially in their sensitivity"(p475) and that "the TOMM misclassified 69% of the claimants who produced inadequate effort on at least one other test."(p479) Although the authors initially considered an argument that "the WMT is overly sensitive and thereby prone to false-positive findings of response bias,"(p480) they subsequently concluded, "failures on the WMT are not the result of excessive sensitivity leading to false-positive determinations"(p480) and "the WMT is a more sensitive measure of response bias than the TOMM."(p485)

We evaluated this assertion by Gervais et al[33] in light of the TOMM TVS we constructed in Frederick and Bowden[5] for healthy adults and patients across 24 studies (FPR = 0.056, TPR = 0.742; see Figure 5). Whether or not they have the same FPRs or TPRs, if the TOMM and WMT are measuring the same construct, and thus directly comparable in the manner conducted by Gervais et al, we can estimate the construct BRs of the Gervais et al samples by use of the TOMM TVS (Figure 5). The

**TABLE 2** *Distribution of VIP Classifications by MMPI-2 VRIN Score*

| VRIN | Compliant | Inconsistent | Irrelevant | Suppression | Total |
|---|---|---|---|---|---|
| 0 | 0.01 | 0 | 0 | 0 | 0.01 |
| 1 | 0.03 | 0.01 | 0 | 0 | 0.02 |
| 2 | 0.07 | 0.01 | 0 | 0.06 | 0.05 |
| 3 | 0.11 | 0.04 | 0.08 | 0.06 | 0.09 |
| 4 | 0.13 | 0.13 | 0 | 0.09 | 0.12 |
| 5 | 0.16 | 0.10 | 0.02 | 0.06 | 0.12 |
| 6 | 0.12 | 0.13 | 0.10 | 0.12 | 0.12 |
| 7 | 0.11 | 0.10 | 0.21 | 0.12 | 0.12 |
| 8 | 0.06 | 0.07 | 0.10 | 0.15 | 0.07 |
| 9 | 0.06 | 0.07 | 0.04 | 0.09 | 0.06 |
| 10 | 0.05 | 0.10 | 0.02 | 0.03 | 0.06 |
| 11 | 0.04 | 0.03 | 0.02 | 0.09 | 0.04 |
| 12+ | 0.05 | 0.21 | 0.40 | 0.12 | 0.13 |
| Total | 296 | 150 | 48 | 33 | 527 |

Abbreviations: MMPI-2, Minnesota Multiphasic Personality Inventory–2; VRIN, Variable Response Inconsistency.

rate of positive TOMM scores in the Gervais et al psychological group was 0.17. In Figure 5, the TOMM TVS, a *y* value of 0.17, corresponds with BR = 0.165. At BR = 0.165, the rate of positive WMT scores = 0.43. The rate of positive TOMM scores in the vocational group (0.01) corresponds to a BR = 0 in the TOMM TVS. The rate of WMT failures is 0.12 at BR = 0. With our estimations of BR (here we construe BR as "the probability that a test taker intends to do poorly on the TOMM" or simply "the probability of malingering") for each of the Gervais et al groups, we now have paired values to construct a TVS for the WMT: (0, 0.12) and (0.165, 0.43). This TVS is presented as Figure 8.

Figure 8 shows that WMT FPR does not remain constant as the probability of malingering increases. The first data point is (0, 0.12). By observation, the WMT FPR is estimated at 0.12 when the probability of malingering is 0. The second data point is (0.165, 0.43). The PPL constructed by joining these points leads to an intersection with the right-hand *y* axis (estimated TPR) that is well in excess of 1. In fact, when the probability of malingering is 0.484, the PPL reaches a maximum (at this point, TPR = 1). This result suggests that when the probability of malingering exceeds 50%, a test taker will routinely, if not always, generate a positive score on the WMT. Clearly, this is an undesirable prospect. It is tantamount to recording "heads" on every flip of a coin simply because the probability of "heads" is 50%.

If we assume WMT TPR = 1 (at BR = 1) and redraw the PPL, using 3 data pairs–(1, 1), (0.165, 0.43), and (0, 0.12)–then the WMT FPR is estimated at 0.20. If the WMT TPR is lower than 1, which seems a realistic assumption, then the FPR estimate of 0.20 will be even higher. The problem is not just that the WMT FPR appears to be high (at least 0.12–0.20) at BR = 0, but it is that the WMT FPR appears to increase dramatically and unrealistically as the probability of malingering increases.

Thus the WMT TVS (Figure 8) contradicts the inference of Gervais et al[33] that the WMT is not "overly sensitive." Whatever constructs positive scores on the WMT actually represent, the WMT TVS demonstrates an unrealistically steep PPL, with a slope greater than 1. Perhaps this is the meaning of "oversensitivity." It may be that the WMT is oversensitive to individuals who have not chosen to malinger, but may be less than fully invested in revealing their maximal abilities. Our interpretation is that the WMT is capturing both "intention" and "effort" and that the use of a dichotomous classification rule (instead of a 4-fold classification scheme) makes it appear oversensitive when compared with the TOMM.

If we construe the WMT as an "intention test" and an "effort test" as those constructs are represented in Figure 7, then Figure 8 is less problematic. In such a construction, the WMT is "overly sensitive" only in the sense that it is capturing something in addition to "intention" as measured by the TOMM, and we believe that this additional construct is "effort." In our view, the primary basis for the apparent nonlinear relationship between "probability of intention to respond correctly" and WMT failure rates, as reflected in Figure 8, is that the WMT, because of its dichotomous classification system, fails to distinguish between the 2 constructs and treats "inconsistent effort" and "bad intention" as the same construct. To further investigate this prospect, in our next analysis, we evaluate a test that uses a 4-fold decision matrix and assesses what happens when we change the decision rules about response style to a dichotomous classification scheme.

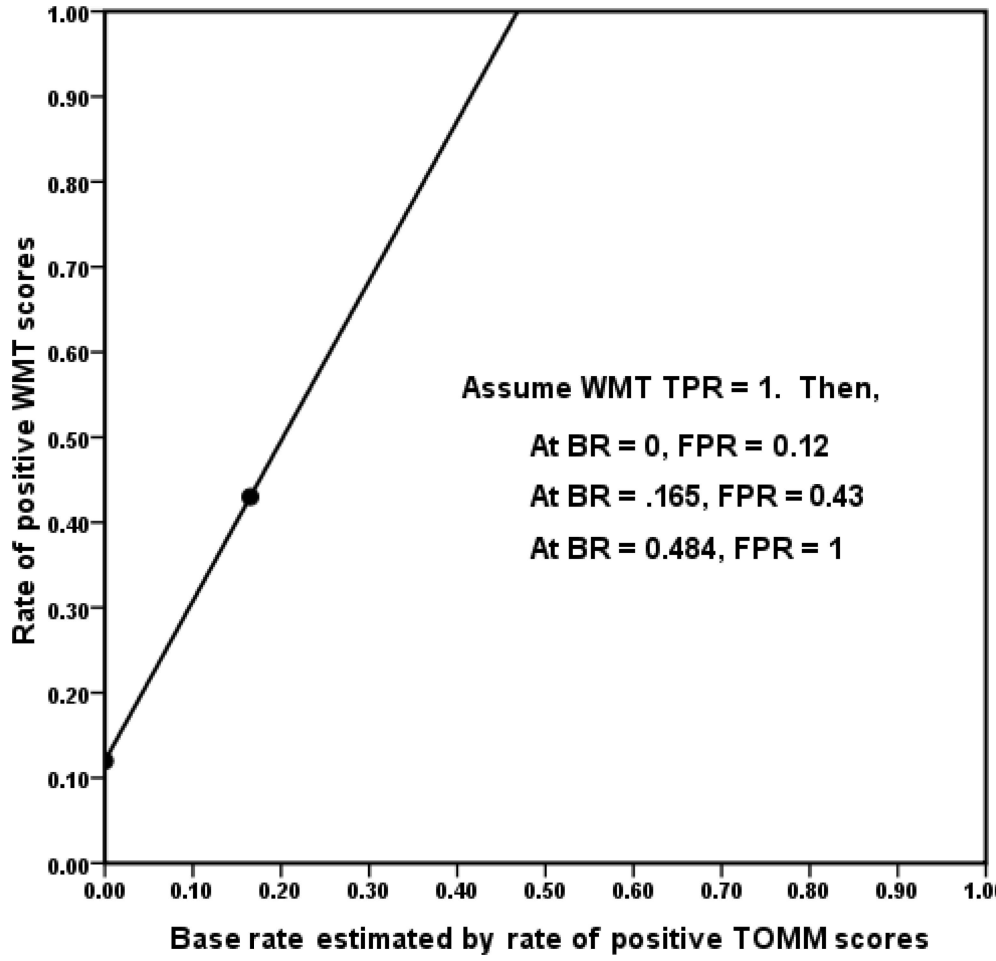


**Figure 8.** TVS for the Word Memory Test. Predictor variable is BR of malingering estimated by rate of positive Test of Memory Malingering (TOMM) scores (failures) in samples reported by Gervais et al.[33] On the basis of TOMM failure rates, and given FPR and TPR estimated for TOMM in Frederick & Bowden,[5] the WMT shows a substantial FPR (0.12) when the BR is estimated at 0, but the FPR increases dramatically as the BR increases. When the BR of malingering is at 0.165, the WMT FPR is at least equal to 0.305, and the WMT FPR is estimated at 1.0 when the BR is as low as 0.484.

## EXPLORING CONSTRUCT VALIDITY OF THE VIP WITH THE TVS VALIDITY INDICATOR PROFILE

"The VIP test is a measure of response validity that is intended to be administered concurrently within a battery of cognitive tests."[18(p1)] The VIP consists of 2 subtests. The Nonverbal subtest includes 100 picture-matrix problems, and the Verbal subtest includes 78 word definition problems. Items within each subtest span a hierarchy of difficulty, from very easy to very difficult. Items are presented in a randomized order of difficulty. Test takers must complete each item, which has 2 answer choices. Once the test is completed, the answers are scored (0 or 1) and the answers are reordered by item difficulty. Spans of 10 scores are averaged, using running means, to yield performance curves comprising 91 (Nonverbal) or 69 (Verbal) points. In a VIP performance curve, the *x* axis of the performance curve represents item difficulty and the *y* axis represents mean performance accuracy. By analysis of the characteristics of the performance curve, and by other measures such as total items correctly answered, the VIP categorizes performance as *Compliant*, *Inconsistent*, *Irrelevant*, or *Suppression*.[18,30,31,34–36] (Figure 7) Frederick et al[31] reported construct validation of these categories, demonstrating that features of the performance curve identified both intention to perform poorly and low effort.

### The VIP and the RMT

To evaluate the viability of our hypothesis that failing to distinguish between 2 constructs can lead to the appearance of oversensitivity, we investigated what constructs the VIP is capturing by using performance on the RMT to generate *x* values in regression analysis. We construe the RMT as an intention test, because effort is a relatively unimportant component of successful

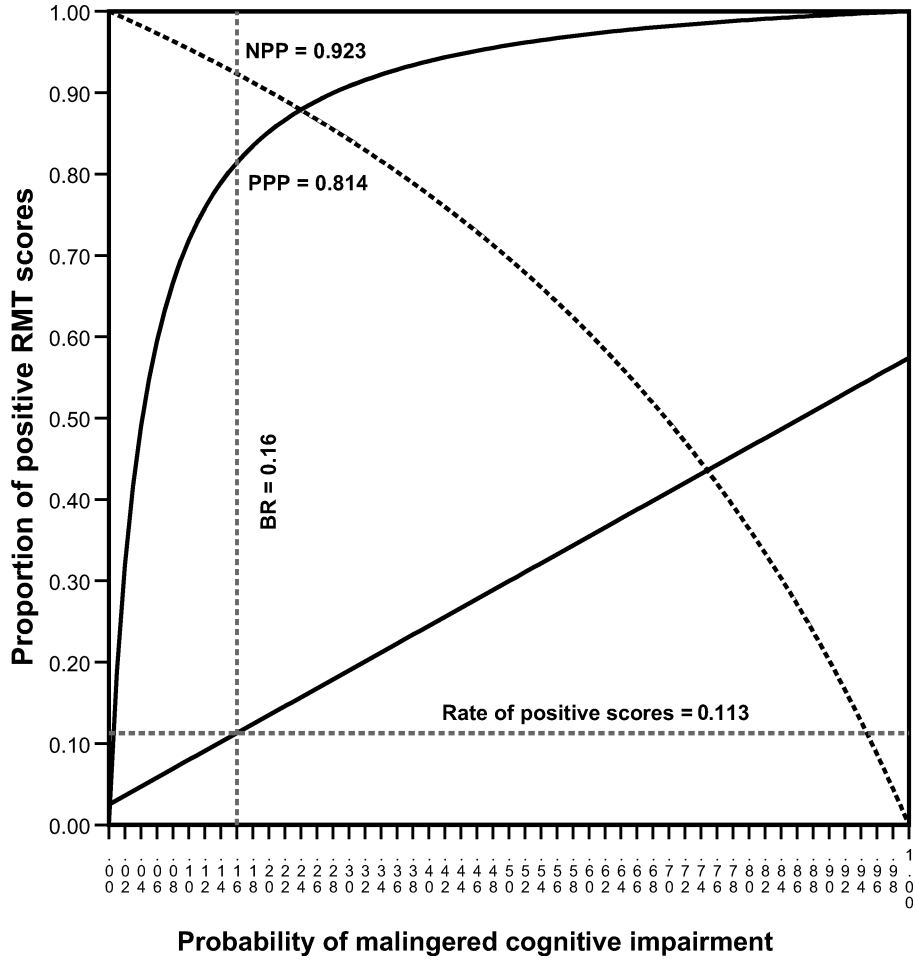


**Figure 9.** TVS for the Rey 15-Item Memory Test (RMT). The figure demonstrates how the RMT TVS can be used to generate predictor variables for validation of another test. In a sample of 731 defendants, the rate of positive RMT scores was 0.113. Using the RMT TVS, this corresponds to a BR of malingering equal to 0.16. At BR = 0.16, the NPP of the RMT is 0.923 and the PPP of the RMT is 0.814. If a person had a negative RMT score, the probability of malingering for that individual was given as 1 − NPP, or 0.077. If a person had a positive RMT score, the probability of malingering was given as PPP, or 0.814. Within subsamples of criminal defendants, the probability of malingering used as the *x* value was the mean value of each individual's probability of malingering based on the RMT score (as in Figure 5).

performance. Therefore, we believe that RMT performance as a predictor variable represents the probability that an individual did not intend to respond correctly to all items on the VIP, as given by performance on the RMT.

Participants were a subset of the sample reported in Frederick[20(pp705–706)] and in our analysis of RMT above. These were 527 participants who completed the VIP, the RMT, and the MMPI-2, and had a clinical rating before testing of the probability they would feign cognitive impairment. All were male criminal defendants undergoing pretrial mental health examination. Mean age was 36.4 (SD = 10.9) years; mean years of education was 10.8 (SD = 3.1). Ethnicity included 51.6% white, 26.6% black, 9.5% Hispanic, 3.8% native American, 0.6% Asian, and 0.6% Pacific Islander. Types of evaluations included competency to stand trial (71.0%), criminal responsibility (43.6%), competency restoration (15.8%), and presentence (8.4%). For simplicity, we used only the VIP Nonverbal categorizations. To have a sufficient number of plotting points, each based on a reasonable number of members, we sorted participants by primary clinician ratings and formed 21 subgroups by consecutive selection; the first 20 subgroups were composed of 25 members; the 21st subgroup had 27 members.

We used the RMT FPR (0.025) and TPR (0.574) derived from ratings of feigned cognitive impairment for criminal defendants (see Figure 6, solid line) to create a TVS for the RMT (Figure 9). The rate of RMT < 9 in the sample was 0.113. From the PPL and the prediction curves in Figure 13, we determined the

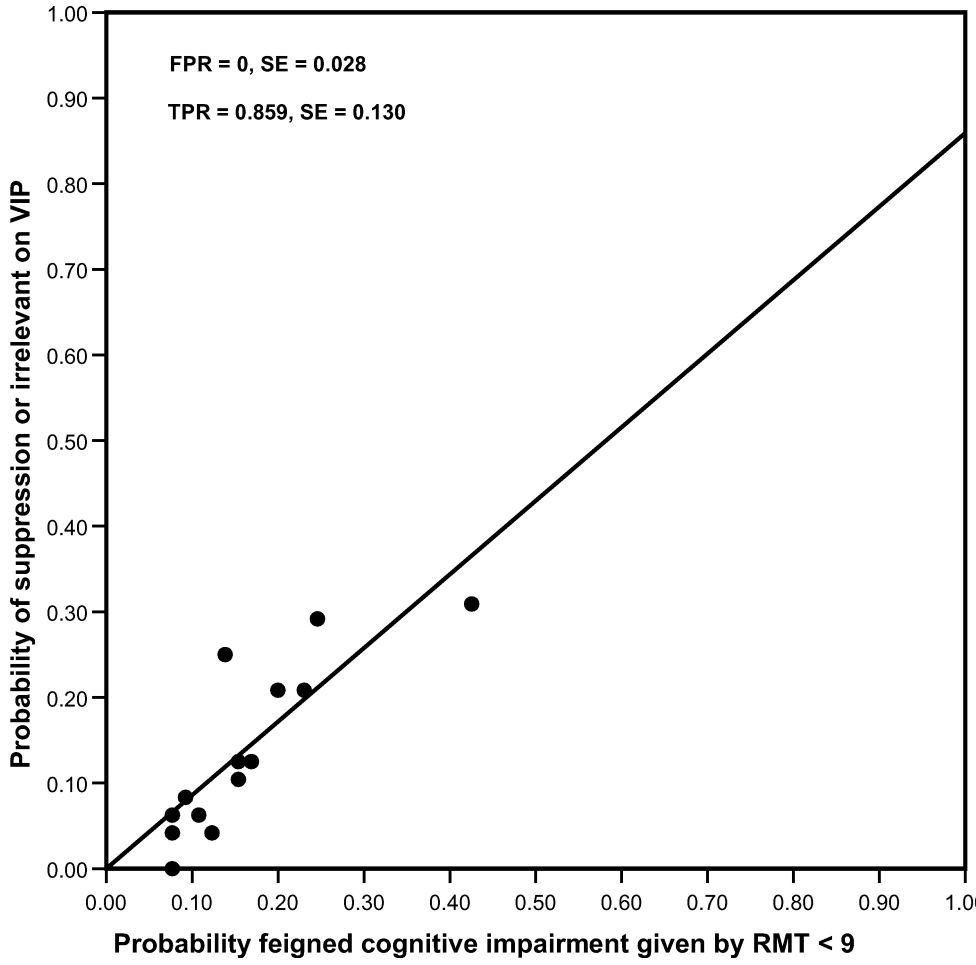


**Figure 10.** Validation of the Validity Indicator Profile (VIP) categories of Irrelevant and Suppression, based on the probability of malingering yielded by Rey 15-Item Memory Test (RMT) score (see Figure 9) for 731 criminal defendants. The categories of Irrelevant and Suppression reflect "no intent to respond correctly," which is the assumed meaning of a failure on the RMT.

following estimates: BR = 0.16, PPP = 0.814, and NPP = 0.923. We therefore assigned each participant a probability of feigned cognitive impairment equal to 0.077 (1 − NPP) if the RMT value was 9 or higher and 0.814 (PPP) if the RMT was 8 or lower. We computed the mean probability value within each subgroup and used these as the *x* values. We computed the proportions of individuals in each subgroup who generated VIP categorizations of Suppression and Irrelevant, the 2 categories representing intention to perform poorly, and used these rates of positive scores as our *y* values. We performed WLS regression analysis on the 21 pairs of *x*, *y* values.

Figure 10 shows the PPL resulting from this analysis. WLS regression results in estimates of FPR = 0, SE = 0.028 and TPR = 0.859, SE = 0.130. If RMT performance captures intent to perform poorly, then the VIP categories that putatively assess bad intent (ie, Irrelevant and Suppression) reveal extremely high sensitivity and specificity. We note the much larger SE for TPR than FPR. This results because the data are positively skewed; there are no mean probability estimates greater than 0.45. Consequently, there are fewer data points to support the estimation of TPR as compared with FPR. In both FPR and TPR, the VIP reveals better classification rates than the RMT, which was used as the basis for validation: FPR (0 vs 0.025); TPR (0.859 vs 0.574). This application of mixed-groups validation[20] (MGV) exemplifies the assertion of Dawes and Meehl[37] that "a psychological test may be validated on a test of lesser validity."[(p65)]

If we continue to use RMT performance as the predictor variable (representing "bad intention"), but now we use the rate of Suppression, Irrelevant, and Inconsistent VIP performance (ie, VIP classification as Invalid) as the criterion variable, we ostensibly change the construct being predicted. We are no longer predicting "bad intention" alone; we are predicting "bad intention and poor

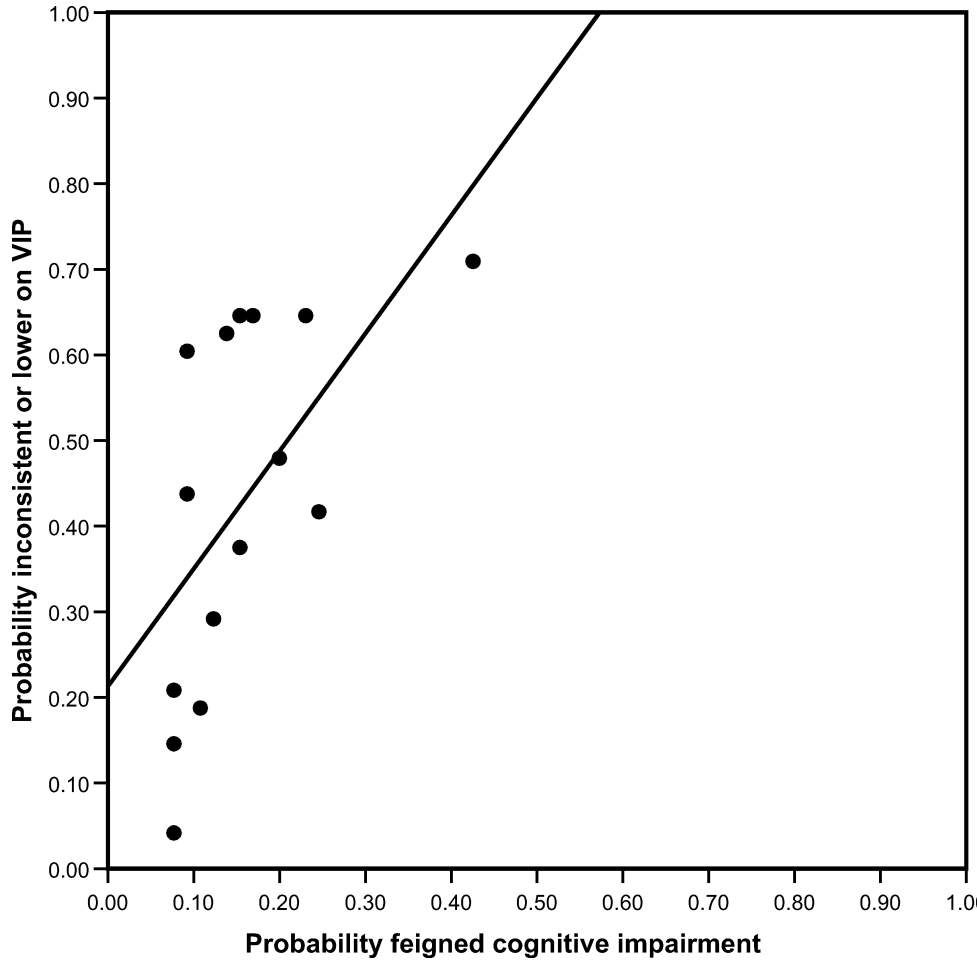


**Figure 11.** Validation of the Validity Indicator Profile (VIP) categories of Inconsistent, Irrelevant, and Suppression on the basis of the probability of malingering yielded by Rey 15-Item Memory Test (RMT) score. Failure on the RMT is assumed to reflect no intent to respond correctly, but the VIP category of Inconsistent assumes that the test taker intends to respond correctly. Therefore, the predictor construct and the criterion constructs do not match. As a consequence, the PPL rises with slope greater than 1, reflecting an increasing false-positive rate as the probability of malingering increases. Given that this was the same outcome in Figure 7, when failure rates on the Test of Memory Malingering were used to validate failure rates on the Word Memory Test, the conclusion is that those 2 tests capture different constructs as well.

effort." This will produce a mismatching of the predictor and criterion constructs.

In Figure 11, we observe what happens when we attempt to validate a construct that is different from the construct represented by the predictor variable–we generate a TVS similar to that observed when the WMT was validated on TOMM outcomes (Figure 8). In Figure 11, we see that as the probability that a person does not intend to respond correctly to all test items increases, the probability of observing a VIP "Invalid" outcome increases dramatically and unrealistically. Consequently, using a dichotomous classification rule for the VIP (ie, Valid vs Invalid) results in the same sort of inflation of FPR we observed for the WMT in Figure 8. The VIP FPR increases in a dramatic and unrealistic manner as the probability of bad intention increases. This supports our hypothesis about the similar FPR inflation for the WMT in Figure 8. When a variable representing intention alone is used to predict the change in a variable representing intention and effort, the test under validation will appear "overly sensitive."

To further evaluate our hypothesis, we generated a predictor variable to more clearly represent "effort" from existing data. To provide a "probability of inconsistent responding," we used the raw score for the Variable Response Inconsistency (VRIN) scale of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2).[38] VRIN is a measure of inconsistent responding, and we believe that efforts can be inferred from consistency of response. When VRIN = 12, the probability of inconsistent responding on the MMPI-2 is generally considered

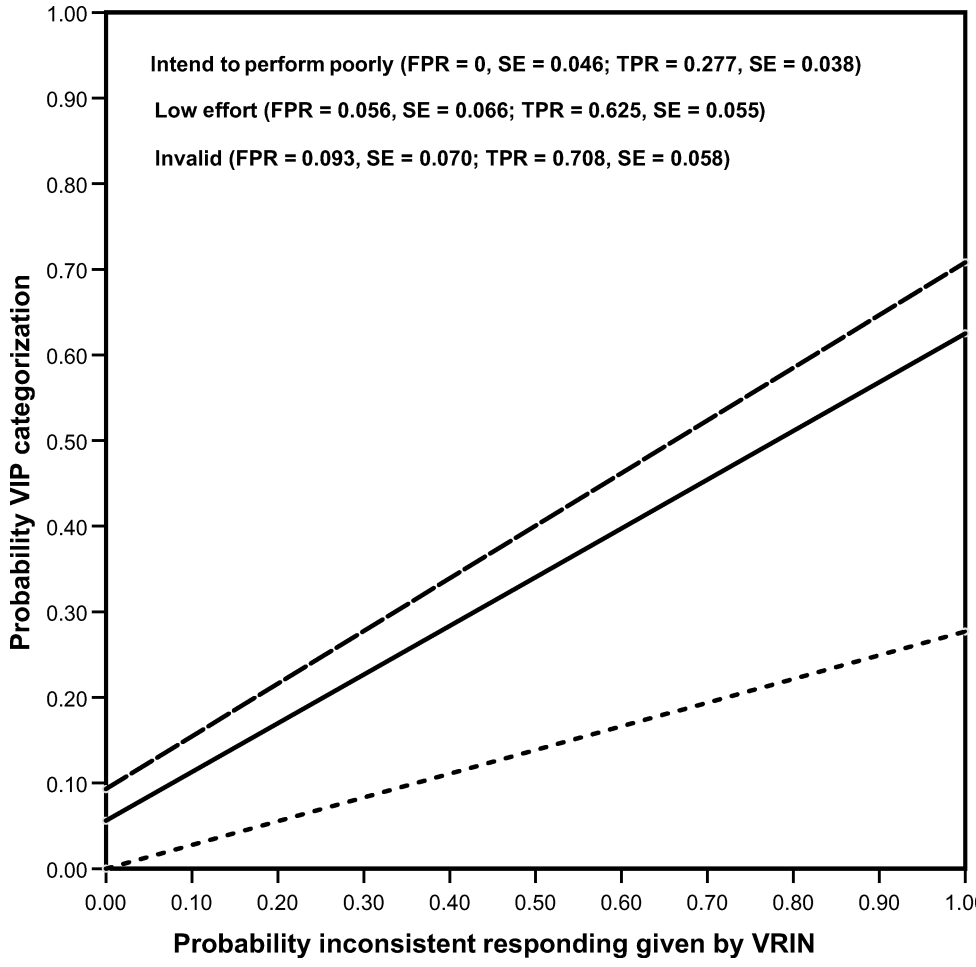


**Figure 12.** Positive proportion lines for 3 configurations of VIP classifications of response style for 527 criminal defendants. The predictor variable is "probability of inconsistency" as estimated by the MMPI-2 VRIN score. VRIN scores were truncated to 12, and all VRIN values were then divided by 12 to obtain a probability estimate. A configuration of Suppression and Irrelevant, (intend to perform poorly, short-dashed line) results in FPR = 0, but TPR = 0.277. Configurations of "low effort" (Inconsistent and Irrelevant; FPR = 0.056 and TPR = 0.625, solid line) and Invalid (Inconsistent, Irrelevant, and Suppression; FPR = 0.093 and TPR = 0.708, long-dashed line) are much more sensitive to the probability of inconsistency. See the text for discussion.

definite.[39] When VRIN = 0, the probability of inconsistent responding is considered quite low. For the purpose of illustration and to approximate probability values between 0 and 1, we, therefore, created a "probability of inconsistent responding." First, we truncated all VRIN > 12 to a value of 12. Then we divided all VRIN by 12, yielding a "probability of inconsistent responding" between 0 and 1. This transformation assumes that the probability of inconsistent responding increases in a linear progression across VRIN raw scores, from 0 to 12. We doubt that this is strictly true, but we argue that there is nonetheless an increasing probability of inconsistent responding as VRIN gets higher. We also doubt that there is a very close correspondence between inconsistent responding on a written personality test and a nonverbal cognitive test. Our long-range goal is to refine methods of estimating probabilities associated with problematic response styles, and we must begin with some less refined methods.

Our immediate goal was to determine whether we could refine the construct we were attempting to validate so that we could understand what might be happening in Figures 8 and 11. In Figure 12, we regressed 3 rates of VIP performance on our estimates of the probability of inconsistent responding:

1. "Intention to perform poorly" rates comprise the rates of Suppression and Irrelevant VIP categories (as represented in Figure 11).
2. "Low effort" rates comprise the rates of Inconsistent and Irrelevant VIP categories.
3. "Invalid" rates comprise the rates of Suppression, Irrelevant, and Inconsistent VIP categories.

Our goal was to determine which of these 3 potential groupings of VIP categorizations would show the closest correspondence to "effort."

We recall that 2 instances of using "intention" to predict "effort" resulted in oversensitivity (Figures 8 and 11). In contrast, in Figure 12, the PPL for VIP "intend to perform poorly" (short-dashed line) represents an instance of insensitivity. Using inconsistent responding as measured by VRIN as a predictor that primarily captures "effort" has an FPR = 0, but a very low TPR (0.277) in the prediction of VIP "intention to perform poorly."

In Figure 12, the PPLs for VIP low effort (solid line) and VIP Invalid (long-dashed line) are almost parallel. They both have a low FPR and moderately high TPR. In considering which of those 2 lines better represents the construct represented by VRIN "probability of inconsistent responding," we chose "low effort." First of all, the 2 VIP categories of "low effort" (Irrelevant and Inconsistent) categorize individuals for which VIP interpretation includes "low effort."[31] This readily relates to the construct we believe that VRIN captures. "Invalid" includes the VIP categorization of Suppression, but Suppression is construed as a category of "high effort."[31] "Invalid" reveals a better TPR than "low effort," but it has a higher FPR.

To evaluate which criterion category better corresponds to the predictor variable, we considered the information in Table 2, which cross-categorizes the rates of VIP classification and VRIN raw score. We noted that classifications of Suppression were distributed fairly evenly across all VRIN raw scores, suggesting that applying strong effort to answer incorrectly on a cognitive test did not correspond to carelessness on a personality test. Consequently, we considered this as support for a decision not to include Suppression in the validation. In contrast, the distributions of VRIN scores are heavily skewed for VIP Inconsistent and Irrelevant categorizations. Therefore, in our view, the validation of the VIP construct of "low effort" (inconsistent-irrelevant responding) is best captured by rates of Irrelevant and Inconsistent VIP categorizations as the criterion variable. Using "probability of inconsistent responding," based on VRIN scores results in FPR = 0.056 (SE = 0.066, 95% CI = 0–0.185) and TPR = 0.625 (SE = 0.055, 95% CI = 0.518–0.732).

It is instructive to consider the difference in how Frederick[18] reported VIP classificatory accuracy and how we have estimated it in these studies. Frederick used ostensibly "pure" criterion groups of Compliant (100 normal persons and 61 neuropsychology patients) and Noncompliant (50 randomly generated protocols, 52 simulators, and 49 patients "at risk for malingering"). He computed FPR = 0.06 and TPR = 0.66 (Table 9),[p35] for any VIP categorization of invalidity (ie, Inconsistent, Irrelevant, or Suppression). Criterion groups validation (CGV) does not facilitate using these participants to validate the construct of "intention to perform poorly" separately from "invalid responding." For example, a CGV validation of "intention" would require looking at the rates of Suppression and Irrelevant responding alone. From his Table 9, this results in FPR = 0.019 and TPR = 0.503. This TPR estimate from CGV is far below the estimate of 0.859 obtained by MGV, and it calls into question the "purity" of the "At Risk for Malingering" and simulating participants used for the VIP validation sample—it is likely that BR < 1 for these subgroups.

Furthermore, CGV analysis of "inconsistency" is not conceivable with these data. Frederick could have estimated the FPR for inconsistency by examining the rate of Inconsistent and Irrelevant classifications in the Compliant group (FPR = 15/161 = 0.093), but there would not be a sensible basis to compute TPR for inconsistency except for the randomly generated protocols (TPR = 49/50 = 0.95). Simulators received instructions to pretend to have memory and concentration problems.[18(p18)] The rate of Inconsistent and Irrelevant responding among these participants was 29/52 or 0.558. We cannot know if this rate reflects bad intention or poor effort.

In contrast, the use of MGV and the TVS has facilitated a more coherent analysis of VIP constructs than possible by CGV. In this study, using criminal defendants undergoing mental health evaluations, a primary target population for application of the VIP, the estimation of VIP classificatory accuracy for identifying "no intention to respond correctly" was FPR = 0, TPR = 0.859. The estimation of VIP classificatory accuracy for "inconsistent responding" was FPR = 0.056, TPR = 0.625. MGV resulted in a much cleaner construct validation and improved classificatory rates when compared with the VIP CGV.

We suggest that these MGV analyses of the VIP with the TVS offer an explanation for the problems presented by Figure 8. We believe that the VIP captures 2 processes, intention and effort. When rates of VIP "invalidity" are regressed on the probability of "bad intention" alone (as in Figure 12), we observe unrealistic FPRs that increase dramatically as the probability of "bad intention" increases. Because the VIP, when used as designed, measures intention and effort as separate variables, it is possible, using MGV and the TVS, to validate these 2 separate constructs with different predictor variables. Because the WMT uses a dichotomous classification rule; however, it cannot distinguish between intention and effort. Clinicians who use tests should have a solid basis for construing what the test is measuring. It seems to us that there is no basis to construe "intention" and "effort" as the same construct. As different constructs, they cannot be measured by a single score. In this sense, we believe that the WMT casts the same pallor over all

problematic test styles, and the FPR associated with its use in this manner is likely quite high.

## DISCUSSION

In this article, our analyses have supported the proposition by Frederick[18,29] that a dichotomous classification scheme within some "malingering" tests results in misclassification of individuals as "malingerers" who do not intend to perform poorly but who may instead present with some sort of compromised effort. The use of the label "effort test" is too loosely applied to the WMT. Consequently, claims of superiority of the WMT with respect to the TOMM[32,33] likely result from an inappropriate comparison of different symptom validity constructs. The TOMM appears to work very well as an "intention" test. When used as an intention test (with result "malingering" vs "nonmalingering"), the WMT likely has an inflated FPR. The WMT may capture both effort and intention but currently does not have a means to quantify those constructs separately.

The TVS and MGV facilitate construct validation of tests. We provided evidence that the WMT[17] should not be directly compared with the TOMM because the 2 tests are likely capturing different constructs related to response style. We noted that the dichotomous classification of the WMT likely makes it oversensitive to aspects of response style by using the term *poor effort* to refer to any aspect of malingering. We showed that if we applied a dichotomous classification scheme to the VIP,[18,29] which uses a 4-fold classification scheme, we could produce the same characteristics of oversensitivity observed in the WMT.

This article extends our initial investigation of the TVS[5] as a means to address problems in reports of test validation studies. The construction of the TVS is based on a straight line drawn between FPR and TPR, the PPL. The PPL communicates the steady, linear increase in the rate of positive test scores as the BR of the condition increases. Because the TVS has the 3 elements necessary to determine predictive power (FPR, TPR, and BR), the TVS also presents graphical representation of NPP and PPP as BR increases from 0 to 1. This alone makes the TVS a worthwhile tool for inclusion in any validation study. Consumers of test validation studies should be able to observe the influence of BR on NPP and PPP based on estimated values of FPR and TPR.

The TVS readily incorporates the methodology of MGV, which provides estimates of FPR and TPR when BRs in a validation sample are not 0 or 1. In this study, we were able to use mixed groups formed by probabilities generated by performances on a test (either the TOMM or the RMT). We showed that probability judgments can also serve as reliable predictor variables in the construction of a TVS. By examination of the resulting diagnostic characteristics, we can evaluate what constructs tests are capturing. Future research should examine how to improve the reliability and validity of probability judgments used as predictor variables.

It is impractical to form "pure" criterion groups for some diagnostic categories that may capture dimensional characteristics. "Effort" is not likely a dichotomous characteristic but likely represents a varying rate of intensity of application of ability in test completion. In our analysis of VRIN and the VIP, we were able to generate probability estimates for "inconsistency" that highlighted the ability of the VIP to capture inconsistency as a response style.

MGV facilitates the use of a less valid test to validate a more valid test. The interested reader is referred to Dawes and Meehl[37] for a detailed exposition of a long-neglected innovation in test validation.[20] In our examination of the VIP, use of the TVS and MGV created a predictor variable based on the RMT (with estimated FPR = 0.025 and TPR = 0.574). The VIP TVS yielded FPR = 0 and TPR = 0.859. In CGV, the estimations of diagnostic characteristics of a test under validation are always limited by the error inherent in the selection criteria. In almost all instances, the FPR and TPR derived by CGV for a test under validation will be worse than those for the selection criteria. In CGV, the selection criteria are assumed to have no error; that is, estimates of BR for the validation samples are restricted to 0 and 1–but this is surely a myth. We know of no perfect selection criteria in psychological research. Consequently, imperfect classificatory rates derived for the test under validation will be ascribed to the new test and not to the limiting factors of CGV.

We believe that the TVS and MGV in combination will best benefit the evolution of test development as a means to (1) display the import of FPR and TPR, (2) provide more accurate and stable estimates of FPR and TPR, (3) report standard errors of estimation for FPR and TPR, (4) provide graphical displays of the behavior of NPP and PPP across BRs, and (5) facilitate construct validation of tests by creative generation of predictor variables without the strictures of CGV. We recommend that the TVS become a standard method for reporting test score diagnostic characteristics and for building upon prior findings from independent validation studies.

## REFERENCES

1. Bush SS, Ruff RM, Troster AI, et al. Symptom validity assessment: practice issues and medical necessity: NAN policy & planning committee. *Arch Clin Neuropsychol*. 2005;20:419–426.

2. Iverson GL, Binder LM. Detecting exaggeration and malingering in neuropsychological assessment. *J Head Trauma Rehabil*. 2000;15(2):829–858.
3. Faust D. Alternatives to four clinical and research traditions in malingering detection. In: Halligan PW, Bass C, Oakley DA, eds. *Malingering and Illness Deception*. Oxford, UK: Oxford University Press; 2003:107–121.
4. Merten T, Bossink L, Schmand B. On the limits of effort testing: symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *J Clin Exp Neuropsychol*. 2007;29:308–318.
5. Frederick RI, Bowden SC. The Test Validation Summary [online]. *Assessment*. Published October 6, 2008. doi: 10.1177/1073191108325005.
6. Tombaugh TN. The Test of Memory Malingering (TOMM): normative data from cognitively intact and cognitively impaired individuals. *Psychol Assess*. 1997;9(3):260–268.
7. Bauer L, O'Bryant SE, Lynch JK, McCaffrey RJ, Fisher JM. Examining the Test of Memory Malingering trial 1 and word memory test immediate recognition as screening tools for insufficient effort. *Assessment*. 2007;14(3):215–222.
8. Lynch WJ. Determination of effort, level, exaggeration, and malingering in neurocognitive assessment. *J Head Trauma Rehabil*. 2004;19(3):277–283.
9. Straus SE, Richardson WS, Glasziou P, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM*. 3rd ed. Edinburgh, UK: Elsevier Churchill-Livingstone; 2005.
10. Wilkinson L. APA task force on statistical inference. Statistical methods in psychology journals: guidelines and explanations. *Am Psychol*. 1999;54:594–604.
11. Green P. The pervasive influence of effort on neuropsychological tests. *Phys Med Rehabil Clin N Am*. 2007;18(1):43–68.
12. Constantinou M, McCaffrey RJ. Using the TOMM for evaluating children's effort to perform optimally on neuropsychological measures. *Child Neuropsychol*. 2004;9:81–90.
13. Ashendorf L, Constantinou M, McCaffrey R. The effects of depression and anxiety on the TOMM in community-dwelling older adults. *Arch Clin Neuropsychol*. 2004;19:125–130.
14. Rey A. *L'examen Clinique en Psychologie* [The psychological examination]. Paris, France: Presses Universitaires de France; 1958.
15. Frederick RI. A review of Rey's strategies for detecting malingered neuropsychological impairment. *J Forensic Neuropsychol*. 2002;2(1):1–25.
16. Lezak MD. *Neuropsychological Assessment*. 3rd ed. New York, NY: Oxford; 1995.
17. Green P, Allen L, Astner K. *The Word Memory Test: A Users Guide to the Oral and Computer Administered Forms, US Version 1.1*. Durham, NC: Cognisyst; 1996.
18. Frederick RI. *Validity Indicator Profile*. 2nd ed. Minnetonka, MN: NCS Pearson Inc; 2003.
19. Rogers R, Harrell EH, Liff CD. Feigning neuropsychological impairment: a critical review of methodological and clinical considerations. *Clin Psychol Rev*. 1993;13:255–274.
20. Frederick RI. Mixed group validation: a method to address the limitations of criterion group validation in research on malingering detection. *Behav Sci Law*. 2000;18:693–718.
21. Cochrane RE, Grisso T, Frederick RI. The relationship between criminal charges, diagnoses, and psycholegal opinions among federal pretrial defendants. *Behav Sci Law*. 2001;19:565–582.
22. Frederick RI. Comment on Faulder Colby: "using the binomial distribution to assess effort: forced-choice testing in neuropsychological settings." *Neurol Rehabil*. 2001;16:309.
23. Green P, Lees-Haley PR, Allen IM. The word memory test and the validity of neuropsychological scores. *J Forensic Neuropsychol*. 2000;2(1):97–124.
24. Green P. Why clinicians often disagree about the validity of test results. *Neurol Rehabil*. 2001;16:231–236.
25. Flaro L, Green P, Robertson E. Word memory test failure 23 times higher in mild brain injury than parents seeking custody: the power of external incentives. *Brain Inj*. 2007;21(4):373–83.
26. Green P. *Green's Word Memory Test: User's Manual*. Edmonton, Canada: Green's Publishing Inc; 2003.
27. O'Bryant SE, Lucas JA. Estimating the predictive value of the Test of Memory Malingering: an illustrative example for clinicians. *Clin Neuropsychol*. 2006;20:533–540.
28. Green P, Iverson G, Allen L. Detecting malingering in head injury litigation with the Word Memory Test. *Brain Inj*. 1999;13:813–819.
29. Frederick RI. *Validity Indicator Profile*. Minnetonka, MN: NCS Assessments; 1997.
30. Frederick RI, Crosby RD. Development and validation of the validity indicator profile. *Law Hum Behav*. 2000;24(1):59–82.
31. Frederick RI, Crosby RD, Wynkoop TF. Performance curve classification of invalid responding on the Validity Indicator Profile. *Arch Clin Neuropsychol*. 2000;15(4):281–300.
32. Green P. Spoiled for choice: making comparisons between forced choice effort tests. In: Boone KB, ed. *Detection of Noncredible Cognitive Performance*. New York, NY: Guilford; 2007.
33. Gervais RO, Rohling ML, Green P, Ford W. A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Arch Clin Neuropsychol*. 2004;19:475–487.
34. Frederick RI. Review of the Validity Indicator Profile. *J Forensic Neuropsychol*. 2002;2(1):125–145.
35. Frederick RI, Foster HG. Multiple measures of malingering on a forced-choice test of cognitive ability. *Psychol Assess*. 1991;3(4):596–602.
36. Frederick RI, Sarfaty SD, Johnston JD, Powel J. Validation of a detector of response bias on a forced-choice test of nonverbal ability. *Neuropsychology*. 1994;8(1):118–125.
37. Dawes RM, Meehl PE. Mixed group validation: a method for determining the validity of diagnostic signs without using criterion groups. *Psychol Bull*. 1966;66:63–67.
38. Butcher JN, Dahlstrom WG, Graham JR, Tellegan A, Kaemmer B. *Manual for Administration and Scoring of the MMPI-2*. Minneapolis, MN: University of Minnesota Press; 1989.
39. Graham JR. *MMPI-2: Assessing Personality and Psychopathology*. 4th ed. New York, NY: Oxford University Press; 2006.