Law and Human Behavior, Vol. 27, No. 3, June 2003 (© 2003)

Daubert, Cognitive Malingering, and Test Accuracy

Douglas Mossman[1,2,3]

Daubert v. Merrell Dow Pharmaceuticals *(1993) held that trial judges should permit expert scientific testimony only when "the reasoning or methodology underlying the testimony is scientifically valid, and . . . properly can be applied to the facts in issue." Vallabhajosula and van Gorp ("V&vG," 2001) have suggested that when the* Daubert *standard is applied to tests for malingered cognitive deficits, courts should deem admissible only results that meet this mathematical standard: assuming a pretest probability of .3, a "positive" score on the malingering test should yield a posttest probability of at least .8. This paper shows that V&vG's criterion may lead to misunderstandings about the kind of information malingering measures provide. After reviewing cases that have discussed both the* Daubert *decision and malingered cognitive deficits, this paper uses data from the Test of Memory Malingering (T. N. Tombaugh, 1996) to provide a general characterization of the mathematical properties of malingering measures. The paper then describes how pretest knowledge about malingering is combined with knowledge about a test's performance to generate a posttest probability of malingering. The results can help mental health experts respond to* Daubert*-inspired challenges to conclusions based on malingering measures.*

KEY WORDS: malingering; Bayes's Theorem; expert testimony; posttest probability; receiver operating characteristic; ROC.

1. INTRODUCTION

At least 60 papers in the PsycINFO database discuss whether tests and diagnostic techniques used by mental health professionals can survive evidentiary challenges to admissibility under the standards promulgated by the U.S. Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals* (1993). In federal courts and in the 30 state jurisdictions (Parry & Drogan, 2000) that have adopted versions of or variations on the *Daubert* standard, a trial judge who faces a "proffer of expert scientific testimony" must undertake "a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or

[1]Division of Forensic Psychiatry, Wright State University School of Medicine, Dayton, Ohio.
[2]University of Dayton School of Law, Dayton, Ohio.
[3]To whom correspondence should be addressed at W.S.U. Department of Psychiatry, P. O. Box 927, Dayton, Ohio 45401-0927; e-mail: douglas.mossman@wright.edu.

0147-7307/03/0600-0229/1 © 2003 American Psychology-Law Society/Division 41 of the American Psychology Association

methodology properly can be applied to the facts in issue" (*Daubert v. Merrell Dow Pharmaceuticals* 1993, pp. 592–593).

Although *Daubert* does not prescribe "a definitive checklist" of matters that judges should consider, the decision lists four factors that could figure importantly in a trial court's assessment of "evidentiary relevance and reliability" (p. 595). These factors are "whether a theory or technique . . . can be (and has been) tested[,] . . . has been subjected to peer review and publication[,] . . . the known or potential rate of error, . . . [and] 'general acceptance'" within the scientific community (pp. 593–594). Although *Daubert* applies specifically to scientific testimony, the Supreme Court has extended federal judges' gatekeeping role to other forms of proffered expert testimony (*Kumho Tire Co. v. Carmichael*, 1999).

As an 1884 U.S. Supreme Court case (*Connecticut Mutual Life Insurance Company v. Lathrop*, 1884) and some pre-1850 state supreme court decisions (e.g., *Clark v. State*, 1843; *McLean v. State*, 1849) demonstrate, U.S. courts have long asked witnesses—including nonprofessionals—to help fact-finders distinguish feigning litigants from those with real mental disorders. Traditionally, mental health professionals have relied on their clinical acumen to detect malingering, noting, for example, whether an evaluee with external (recognizable, contextual) motivation to malinger has produced symptoms or symptom combinations that are "inconsistent with physiological or anatomical mechanisms" (Cunnien, 1997, p. 46) or that are improbable, unusual, and/or contradictory (Resnick, 1997a, 1997b). In the last two decades, however, investigators have developed several measures—structured interviews, questionnaires, self-report instruments, and evaluation exercises—specifically designed to sort honest from feigning evaluees (Pankrantz & Binder, 1997; Rogers, 1997a). Given the now-recognized frequency with which litigants fake or exaggerate mental illness and cognitive deficits (Rogers, 1997b), it is not surprising that several of these malingering measures have found their way into court through reports or testimony of who experts who relied on them.

Recently, Vallabhajosula and van Gorp ("V&vG," 2001) have asked whether three often-used tests of malingered cognitive deficits—the Rey Fifteen-Item Test ("FIT"; Rey, 1964), the Test of Memory Malingering ("TOMM"; Rees, Tombaugh, Gansler, & Moczynski, 1998; Tombaugh, 1996, 1997), and the Validity Indicator Profile ("VIP"; Frederick & Crosby, 2000)—"likely meet the *Daubert* standard for admissibility of scientific evidence" (p. 207). V&vG examine each test's ability to meet the following Bayesian criterion: a "positive" result on the test must yield a positive predictive value (PPV) of greater than .8, assuming that the pretest probability of malingering equaled .3. V&vG acknowledge that this criterion is "arbitrary," but believe it provides a reasonable "common standard by which to compare . . . the tests" (p. 210). V&vG conclude that TOMM "shows considerable promise," that VIP should be used for now with "some caution," and that FIT fails (p. 214).

This paper does not agree or disagree with V&vG's conclusions, but instead argues that their criterion for admissibility may lead to misunderstandings of the kind of information malingering measures provide. The next section reviews cases that have discussed both the *Daubert* case and the possibility of malingered cognitive deficits, to see whether courts have, in fact, developed any sort of statistical criterion by which to judge tests of malingering measures. Section 3 offers a detailed

description of V&vG's criterion. Section 4 explains why that criterion may generate misunderstandings of the properties of malingering measures, and, using TOMM as an example, provides a better general characterization of the mathematical properties of malingering measures. Section 5 draws on recent medical and psychological publications to show how one combines pretest knowledge about malingering with knowledge about a test's performance characteristics to generate a posttest probability of malingering. Section 6 explains how this paper's calculations and findings might help mental health experts respond to *Daubert*-inspired challenges to conclusions derived in part from malingering measures.

2. CASE LAW AND TESTS OF MALINGERED COGNITIVE DEFICITS

A search of the LEXIS "Combine Federal and State Case Law—U.S." database[4] shows that, through December 2002, 18 published cases refer by name to specific tests of cognitive malingering. Eleven of the cases refer to FIT, two cases mention TOMM, two mention VIP, one mentions the modified Hiscock Forced-Choice Procedure (Guilmette, Hart, Guiliano, & Leininger, 1994), and two cases refer to "forced-choice tests" without specifying which instruments were used. The case citations appear in Table 1.

Only four of these cases evaluate malingering measures from the perspective of *Daubert* or other legal rules governing admissibility of evidence. Two cases are 1997 New York federal district court decisions by Judge Jack B. Weinstein concerning the trial and sentencing of Vincent Gigante. The first decision, *United States v. Gigante* (1997a; hereinafter "*Gigante I*"), concerns the defendant's postconviction request for a new trial based on purported "new evidence" that he had been incompetent when tried and convicted. At a hearing on the motion, a defense expert had testified that he had administered "a battery of interactive, neuropsychological tests, including the Portland Digit Recognition Test, the Warrington Face Recognition Test, the Warrington Word Recognition Test, and the Rey 15-Word Test, that were designed [sic] to ascertain whether [the] defendant was malingering." The results of these tests led the expert to conclude "that [the] defendant was not feigning incapacity, but in fact suffered from severe cognitive impairment" (*Gigante I*, p. 147). Judge Weinstein determined, however, that the expert's conclusions

> were based upon speculative assumptions. For instance, he [the expert] argued that since defendant seemed to be "trying" on the cognitive tests and scored slightly above the score of "chance" (that is, the score one might achieve by simply guessing), the conclusion was valid that malingering was not probable.
>
> None of the defendant's experts took into account the extensive testimony received by Judge Nickerson [the previous judge in the case] proving that defendant's mental difficulties had been feigned for many years. . . .

[4]This database includes decisions by the U.S. Supreme Court, all U.S. Courts of Appeals, all federal district courts, and case law for all 50 states plus the District of Columbia, and all U.S. territories. The search strategy—(TOMM AND MALING!) OR "TEST OF MEMORY MALINGERING" OR (REY AND MALING!) OR (((FIFTEEN PRE/1 ITEM) AND TEST) AND MALING!) OR ((HISCOCK OR "FORCED CHOICE" OR "SYMPTOM VALIDITY" OR SVT) AND MALING!) OR (VIP AND MALING!) OR "VALIDITY INDICATOR PROFILE")—was designed to detect all published cases that discuss the tests that are commonly used to detect malingered cognitive deficits. The search was updated February 10, 2003.

Table 1. Published Cases Through December 2002 Referring to Instruments for Detecting Cognitive Malingering

Rey 15-Item Test
United States v. Soldevila-Lopez, 17 F.3d 480, (1st Cir. 1994)
Harris v. Vasquez, 913 F.2d 606 (9th Cir. 1990)
Harris v. Vasquez, 949 F.2d 1497 (9th Cir. 1990), *mandate stayed*, 1991 U.S. App. LEXIS 33729 (November 15, 1991), *reh'g denied and suggestion for reh'g en banc rejected*, 1991 U.S. App. LEXIS 29419 (November 8, 1991)
Downs v. Perstorp Components, Inc., 126 F. Supp. 2d 1090 (E.D. Tenn. 1999)
Hinds v. Apfel, 1999 U.S. Dist. LEXIS 20197 (N.D. Calif., December 3, 1999)
Ingram v. Martin Marietta, 36 F. Supp. 2d 1190 (C.D. Calif. 1999)
United States v. Gigante, 982 F. Supp. 140 (E.D. N.Y. 1997)
United States v. Gigante, 996 F. Supp. 194 (E.D. N.Y. 1997)
Smetana v. State, 991 S.W.2d 42 (Tex. App. 1998)
Morris v. Barnhart, 2002 U.S. Dist. LEXIS 16933 (E. D. Mich., June 28, 2002)
Harbison v. State, 2002 Tenn. Crim. App. LEXIS 315 (February 13, 2002)
Test of Memory Malingering
Villalba v. Consolidated Freightways Corp., 2000 U.S. Dist. LEXIS 11773 (N.D. Ill., August 10, 2000)
Rez v. City of Los Angeles, 2002 Cal. App. Unpub. LEXIS 999, (Cal. App. 2nd Dist., May 9, 2002)
Validity Indicator Profile
United States v. Walker, (S.D.N.Y., May 5, 1998)
Weishaar v. Barnhart, 2002 U.S. Dist. LEXIS 14175, (N.D. Iowa, February 23, 2002)
Modified Hiscock-Forced Choice Procedure
State v. Finley, 1998 Ohio App. LEXIS 2693, (Ohio App., June 19, 1998)
Unspecified Forced Choice
Marks v. 84 Lumber Co., 771 So. 2d 751 (La. App. 2000)
United States v. Greer, 158 F.3d 228 (5th Cir. 1998), *cert. denied*, 1999 U.S. LEXIS 1664 (March 1, 1999)

> The opinions of defendant's experts were unreliable. They were not consistent with other evidence in the case. Assuming for the sake of this memorandum that their opinions met the *Daubert* test, they were not credible and were unpersuasive. (*Gigante I*, pp. 147–148)

The second decision by Judge Weinstein, *United States v. Gigante* (1997b; hereinafter *"Gigante II"*), addresses defense lawyers' postconviction assertion that their client was incompetent to be sentenced. *Gigante II* discusses the findings and opinions of more than a dozen mental health professionals, and includes a lengthy description of results from neuropsychological tests and malingering measures administered by the same defense expert who testified in *Gigante I* (*Gigante II*, pp. 212–216). *Gigante II* contrasts this expert's interpretation of test results with the interpretations of a government expert (pp. 222–227), and finds the latter's interpretations much more persuasive. *Gigante II* does not comment on or evaluate specific aspects of the evidence, which included brain scans as well as tests of malingering. Instead, just one sentence in the decision refers to the *Daubert* standard; the standard is applied globally to the evidence presented by all experts, with whose credentials and qualifications Judge Weinstein was clearly impressed:

> All of the experts assisting the court in dealing with the difficult questions surrounding diagnosis of defendant's mental condition were able, ethical, and candid. They all met the standards of *Daubert*. . . . That the court did not credit the conclusions of those experts tendered by the defendant constitutes no suggestion of lack of confidence in their professional skills and veracity. (*Gigante II*, p. 199)

In the third case, *Downs v. Perstorp Components and ICI Americas* (1999), the two defendant companies sought to exclude expert testimony that the plaintiff suffered a "chemical encephalopathy" following work-related exposure to a compound called Rubiflex. The plaintiff's expert had concluded that the plaintiff was not malingering on the basis of information subtest scores on the Wechsler Adult Intelligence Scale (WAIS) and the plaintiff's good score on FIT. A defense expert testified, however, that WAIS "had never been demonstrated to have anything to do with the measurement of malingering," and FIT was "well known in the neuropsychology literature to be insensitive to malingering" (p. 1109). The district court ultimately excluded testimony by the plaintiff's expert, finding that his views were "essentially based upon his determination, without any scientific basis, that all injuries which occur after exposure to a chemical compound must be causally related to and result from the individual's exposure to chemicals" (p. 1128).

The fourth case, *Villalba v. Consolidated Freightways Corp.* (2000), addresses, inter alia, a plaintiff's motion in limine to exclude testimony by a defense expert that the plaintiff who sought recovery for alleged damages stemming from an automobile and truck collision was faking cognitive deficits. The motion asserted that (1) the expert had not revealed "the norms which apply to the Computerized Assessment of Response Bias (CARB) and TOMM tests or the bases for the comparative figures relied upon to draw his conclusions," and (2) the expert's alleged refusal to answer questions about the tests "means that his opinions will go effectively unchallenged at trial without any showing that his opinions are valid or have been subjected to peer review or meet in any way the requirements set forth by the Supreme Court [in *Daubert*] for controlling the rendering of unreliable opinions from experts" (*Villalba v. Consolidated Freightways Corp.*, 2000, p. 30).

The Court held, however, that the expert's report had met the requirements of Federal Rule of Civil Procedure 26(a)(2)(B), which states that an expert's report should explain the "basis and reasons" for the opinions and "the data or other information" used to reach the opinions. For example, the expert's report said the plaintiff "concocted an error rate [on the TOMM] that would not be seen in any legitimate form of mild head trauma" (*Villalba v. Consolidated Freightways Corp.*, 2000, pp. 31–32); the report also contained charts that compared the plaintiff's scores with "the norms" for the tests. Concerning the second claim, the Court ruled that the plaintiff's "cursory challenge" failed because she had "not provided the Court with any information concerning how to evaluate a neuropsychologist's opinions within the context of the four indices of reliability identified by the Supreme Court in *Daubert* (testing, peer review, error rate, and general acceptance)" (*Villalba v. Consolidated Freightways Corp.*, 2000, p. 30).

Although *Coe v. Tennessee* (2000) does not discuss measures of *cognitive* malingering, it offers an instructive view about how courts might respond to efforts to use *Daubert* as a bar to evidence of malingering. *Coe* concerns a condemned prisoner's appeal of a trial court's determination that he was competent to be executed. At the competence hearing, one of the state's experts had testified that the prisoner was malingering, basing this opinion in on the prisoner's performance on the Millon™ Clinical Multiaxial Inventory, the Structured Interview of Reported Symptoms, and the Minnesota Multiphasic Personality Inventory. The appellant prisoner's lawyers

argued that these tests were an inadequate scientific basis for the expert's conclusion because "(1) they have not been subjected to peer review for their validity in determining malingering in death row inmates; (2) they have not been tested to determine their scientific validity for determining malingering in death row inmates; and (3) the rate of error for these tests has not been established for death row populations" (*Coe v. Tennessee*, 2000, p. 225).

In considering this argument, the Tennessee Supreme Court noted that its jurisdiction's standard of admissibility, though narrower than the *Daubert* standard, utilized "general principles" (*Coe v. Tennessee*, 2000, p. 226) similar to *Daubert* and Fed. R. Evid.702. Under Tennessee's standard, admission of the test results by the trial court was reasonable:

> We find that the psychological tests in question were relevant to a determination of the appellant's competency to be executed. . . . [T]hese types of standardized tests have long been recognized as scientifically valid and reliable. . . . Indeed, the reliability of these tests is further illustrated by the fact that the appellant's rebuttal expert witness . . . had administered almost exactly the same battery of tests as those administered by [the state's expert].
>
> . . . In light of the conflicting scientific views between [the expert witness], the critical inquiry here is not the admissibility, but the weight to be given to [state expert]'s testimony. (*Coe v. Tennessee*, 2000, p. 227)

Five cases are a very small data set from which to draw conclusions about how courts might apply *Daubert* (or similar rules governing admissibility) to malingering measures. One feature that unites these cases, however, is the courts' failure to examine the accuracy (or, more broadly, the reliability) of the malingering measures under consideration. Although a test's error rate is listed explicitly in *Daubert* as one factor a judge should consider before allowing presentations based on that test, none of the above decisions examine this matter. This is noteworthy, given the well-known inadequacies of FIT in distinguishing between honest responders and malingerers (Greiffenstein, Baker, & Gola, 1996; Guilmette et al., 1994; Schretlen, Brandt, Krafft, & van Gorp 1991). The V&vG proposal, which implies that courts should establish a minimum level of accuracy for admitting evidence, thus appears to address an important gap in courts' reasoning about tests of cognitive malingering. Let us now examine that proposal in more detail.

3. CHARACTERIZING THE V&vG PROPOSAL

The V&vG proposal assumes that a malingering test is binary; that is, it will have (or will be used as though it had) just two possible outcomes, "positive" or "negative." Denote these outcomes as $T+$ and $T-$, respectively; denote the presence and absence of malingering as $M+$ and $M-$; denote the prior probability of malingering—the "base rate"—as $\mathrm{P}(M+)$. Given a positive test result, Bayes's Theorem says that the posterior probability or malingering, or PPV of the test, is (Bayes, 1763)

$$\mathrm{PPV} = \mathrm{P}(M+|T+) = \frac{\mathrm{P}(T+|M+)\mathrm{P}(M+)}{\mathrm{P}(T+|M+)\mathrm{P}(M+) + [1 - \mathrm{P}(T-|M-)][1 - \mathrm{P}(M+)]} \quad (1)$$

Here, P($M+|T+$) represents the conditional probability of malingering, given a "positive" test outcome; P($T+|M+$) represents the true positive rate (TPR) or the *sensitivity* of the test, that is, the probability of getting a positive test result if someone is malingering; and P($T-|M-$) represents the true negative rate or the *specificity* of the test, that is, the probability of getting a negative test result if someone is not malingering. Note that a test's false positive rate (FPR) equals 1 − specificity.

V&vG interpret the *Daubert* standard as requiring that courts accept evidence derived from only those malingering measures for which "PPV ≥80 percent with a base rate of ≤30 percent" (V&vG, 2001, p. 211). Using this paper's symbols, V&vG would require that P($M+|T+$) be at least .8 when P($M+$) is .3, or

$$\frac{(.3)\mathrm{P}(T+|M+)}{(.3)\mathrm{P}(T+|M+) + [1 - \mathrm{P}(T-|M-)][1 - .3]} \geq .8 \tag{2}$$

Inspection of Eq. 2 reveals that an infinite number of combinations of sensitivity and specificity will yield an adequately accurate test under the V&vG standard. But these combinations may be conveniently summarized. If one divides both sides of Eq. 1 by 1 − P($M+|T+$), substitutes the right side of Eq. 1 for P($M+|T+$) on the right side of the resulting equation, does some algebra, and then rearranges terms, one gets this result:

$$\frac{\mathrm{P}(M+|T+)}{1 - \mathrm{P}(M+|T+)} = \frac{\mathrm{P}(T+|M+)}{1 - \mathrm{P}(T-|M-)} \times \frac{\mathrm{P}(M+)}{1 - \mathrm{P}(M+)} \tag{3}$$

The left side of Eq. 3 equals what one might call the "positive posterior odds," that is, the posttest odds that an evaluee is malingering, given the a positive test result and pretest probability P($M+$); denote this as O($M+|T+$). The factor on the right side of the multiplication sign equals O($M+$) or the *prior odds*—the odds, before test results were known, that the evaluee is malingering.[5] The factor between the equal sign and the multiplication sign is $LR+$, the *likelihood ratio* associated with a positive test result. Sometimes the properties of diagnostic tests and detection methods are described using the terms *hit rate* and *false alarm rate*. A test's hit rate is its TPR; a test's false alarm rate is its FPR. Thus, for a binary diagnostic test, $LR+$ = sensitivity/(1 − specificity) = TPR/FPR. One can rewrite Eq. 3 as

$$\mathrm{O}(M+|T+) = \frac{TPR}{FPR} \times \frac{\mathrm{P}(M+)}{1 - \mathrm{P}(M+)} = LR+ \times \mathrm{O}(M+) \tag{4}$$

If, according to the V&vG proposal, PPV must be at least .8 when P($M+$) = .3, then $LR+$ must be large enough so that O($M+|T+$) will be at least 4 when O($M+$) = 3/7. Plugging these numbers into Eq. 4 and solving for $LR+$, we see that any test for which $LR+ \geq 28/3$ (or TPR/FPR ≥ 28/3) will satisfy the V&vG criterion.

[5]The relationship between the probability of x, P(x), and the odds of x, O(x), is expressed as O(x) = P(x)/1 − P(x). If a weather forecast asserts, "the probability of rain today is 40%," then odds of rain is 2/3 = 2:3 = 1:1.5.

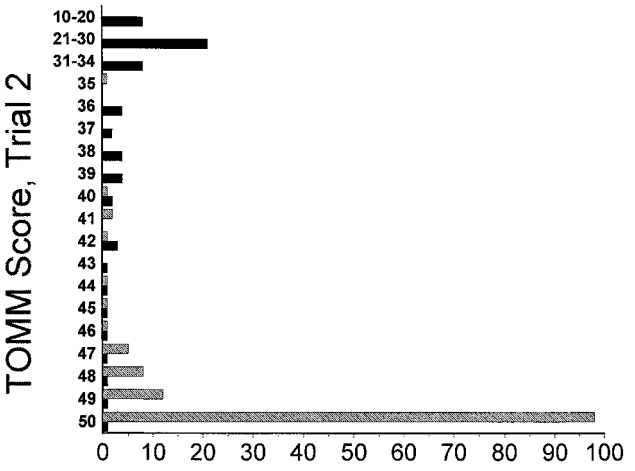

Fig. 1. Combined results from studies of the TOMM, Trial 2. Hatched: honest respondents. Black: simulated malingerers.

4. PROBLEMS WITH THE V&vG CRITERION

The V&vG criterion has five limitations.

Malingering Measures Have More Than Two Outcomes

V&vG's criterion implicitly suggests that a malingering test has only two outcomes—"positive" or "negative." Although malingering tests are often *used* as though they were binary, they usually have several possible outcomes. For example, because evaluees can score from 0 to 50 on TOMM, TOMM has 51 possible outcomes. If readers examine Fig. 1, which contains combined results from several validation studies of TOMM (Rees et al., 1998; Tombaugh, 1996, 1997), they will see two overlapping distributions: a set of higher scores obtained by 131 nondemented, honestly responding persons with genuine neuropsychological problems, and a set of lower scores obtained by 64 persons known to be faking cognitive deficits.[6] Only by

[6]The results shown in Fig. 1 were derived as follows.

One hundred and thirty-one *honest responders with genuine neuropsychological problems*. Appendix A of Tombaugh (1996) gives the actual Trial 2 scores obtained by 108 honestly responding participants who had a variety of cognitive impairments, aphasias, and traumatic brain injuries (TBIs), as well as clinical information about these participants' difficulties (for a summary, see Tombaugh, 1997, pp. 263–264). Information in the text and Table 3 of Rees et al. (1998) allows one to reconstruct the Trial 2 scores of 10 honest responders with mild TBIs, and similar information concerning Table 4 permits reconstruction of Trial 2 scores for 13 nonlitigating participants with TBIs. Although these 131 participants comprise a diagnostically heterogeneous group, their test performances on Trial 2 of the TOMM were very similar (mean scores $\geq$ 48.5 out of 50). Not included in this set of results are findings from participants with dementias, who did comparatively poorly (mean score = 45.7).

Sixty-four *malingerers*. Dr Tombaugh kindly supplied the actual Trial 2 scores obtained by 64 volunteers who were asked to simulate the behavior of head-injured persons. The results shown were obtained from three different testing conditions, but turned out to be similar: for Condition 1, $N = 31$, mean $\pm$ SD = 29.6 ± 8.6; for Condition 2, $N = 25$, mean $\pm$ $SD = 35.7 \pm 8.9$; for Condition 3, $N = 8$, mean $\pm$ SD = 32.1 ± 7.3. These results were previously reported by Tombaugh (1997) and Rees et al. (1998).

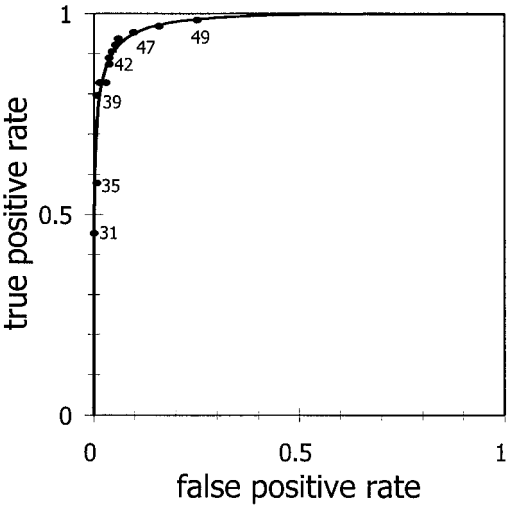

Fig. 2. ROC graph depicting the performing of the TOMM, Trial 2, as its cut-off is varied from 30 to 50.

choosing a particular cutoff to distinguish "positive" results from "negative" results can one say that the test has a particular sensitivity and specificity. For example, if a Trial 2 score below 45 is deemed a "positive" test result, TOMM's sensitivity in detecting malingering is $58/64 = .91$, and its specificity is $125/131 = .95$. If one chooses a higher cutoff (say, 48) the test's sensitivity rises (to $61/64 = .95$), but its specificity falls (to $118/131 = .90$); by lowering the cutoff (to, say, 42), one lowers the sensitivity (to $53/64 = .83$), but raises the specificity (to $127/131 = .97$). This means that the TOMM's accuracy—and the accuracy of any discrimination tool where test results form two overlapping distributions—should be described in a way that conveys the trade-offs between sensitivity and specificity that occur as the cutoff is altered.

Mental health professionals have recognized that receiver operating characteristic (ROC) graphs provide a readily apprehended means of portraying this feature of a detection system's performance, whether that system is used for discerning violence (Douglas, Ogloff, Nicholls, & Grant, 1999; Gardner, Lidz, Mulvey, & Shaw, 1996; Mossman, 1994a, 1994b; Rice & Harris, 1995) malingering (Mossman, 2000a; Mossman & Hart, 1996), or other diagnostic categories (Swets, Dawes, & Monahan, 2000). In a ROC graph, it is customary to plot a detection system's true positive rate as a function of the system's FPR. The result is a succinct pictorial summary of the test's performance throughout its entire range of possible outcomes.

Figure 2 contains a ROC graph that depicts the performance of TOMM as its cutoff is varied from 30 to 50.[7] Notice that in this figure, individual cutoffs are fitted to a smooth ROC *curve*. The curve utilizes the "binormal" assumption of ROC analysis, which suggests that following a monotonic transformation of the original decision axis (here, the TOMM scores), the raw data will fall into two overlapping, Gaussian

[7]None of the honest responders with genuine problems scored below 36.

("bell-shaped") distributions with different means and variances.[8] This means that one can summarize the TOMM test results with the equation

$$z(TPR_i) = A + Bz(FPR_i) \tag{5}$$

In Eq. 5, A equals the difference in means between the two populations, measured in units of the standard deviation of the malingering ($M+$) population, B equals the ratio of test result standard deviations (i.e., SD_{M-}/SD_{M+}), and $z(\text{TPR}_i)$ and $z(\text{FPR}_i)$ are the normal deviates of the test's true and false positive rates at a given score i. In other words, when (FPR, TPR) pairs are plotted as z-scores, the resulting points tend to fall along a straight line. Applying Metz and colleagues' software to the data shown in Fig. 1, one finds that $A = 2.8064$ and $B = 0.9015$.

Distinguishing Between Moderately Accurate Tests and Very Good Ones

A second limitation of the V&vG criterion is that, although it purports to set a high accuracy threshold for admissibility, it does not differentiate between very accurate malingering measures and those that are intrinsically less accurate. Figure 3 illustrates this. There, three ROC curves appear; they depict the performance of three hypothetical tests for detecting malingering. All three ROC curves are constructed by letting $B = 1$, but the values of A equal 1.2, 2.1, and 3. Also shown in Fig. 3 is a diagonal line for which TPR/FPR = 28/3. Any point to left of this line represents a sensitivity–specificity pair that satisfies the V&vG criterion. Because portions of all three curves contain fall to the left of the diagonal line, all three malingering tests have cutoffs that would satisfy the V&vG criterion.

Yet the tests have very different performance characteristics. To appreciate this, suppose that an examiner chose, as each test's operating point, the score at which FPR $\leq$ .02. These cutoffs are represented by ovals on each ROC curve in Fig. 3. Notice that for the test represented by the lowest ROC curve (for which $A = 1.2$), TPR = .197 when FPR = .02, and $LR+ = 9.83$. Returning to Eq. 4, this means that a "positive" result increases the odds of malingering by a factor of 9.83; if the prior probability of malingering is .30, the posttest probability is .81. But for the middle ROC curve, TPR = .518 when FPR = .02, so that $LR+ = 25.9$, and for the upper ROC curve, TPR = .828 when FPR = .02, so that $LR+ = 41.4$. For these two tests, a "positive" result raises the posttest probability of malingering from .30 to .92 for the middle curve, and to .95 for the upper curve. In other words, all three tests would satisfy the V&vG criterion, but the latter two tests justify a higher level of confidence than does the first. To put this another way, at the same FPR the latter two tests—by virtue of their higher TPRs—detect a much larger fraction of the malingerers.

Notice that the more accurate the test, the larger the *area* under the ROC curve (AUC). This is the case because, at any given value of FPR, TPR is larger for the more accurate test, and thus the integrated AUC for the better test is larger than the

[8]For more detailed discussions of ROC curves and the binormal curve fitting assumptions as they apply to forensic mental health issues, see Mossman (1994a, 1994b). Concerning the robustness of the binormal assumptions, see Hanley (1988, 1996). For recent discussions of the limitations of this model, see Dorfman et al. (1997) and Metz and Pan (1999).

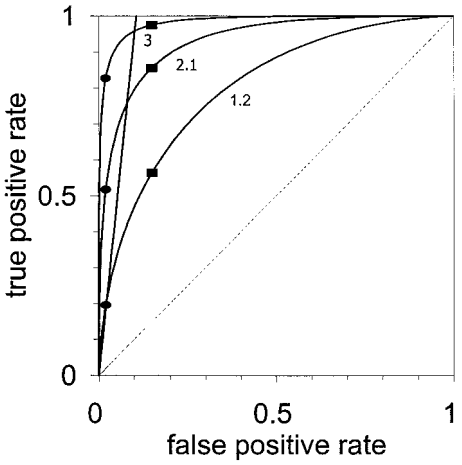

Fig. 3. Three hypothetical ROC curves illustrating relationships among cut-offs, test accuracy, and the V & vG criterion.

AUC for the poorer test. AUC is related to indices A and B (Swets & Pickett, 1982; Somoza & Mossman, 1991a) by the following formula:

$$AUC_z = \Phi\left(\frac{A}{\sqrt{1+B^2}}\right) \tag{6}$$

In Eq. 6, $\Phi(x)$ is the cumulative normal distribution function of x, and the subscript z implies that AUC is calculated from a fitted binormal ROC curve.[9] In this context, AUC also equals the likelihood that, if two individuals are drawn randomly from the $M+$ and $M-$ populations, the malingering measure will assign a score more indicative of malingering to the $M+$ individual (Hanley and McNeil, 1982). For the three ROC curves shown in Fig. 3, AUCs are .802, .931, and .983. Although all three tests can be implemented in a way that satisfies the V&vG criterion, AUC captures important differences that the V&vG criterion ignores. Incidentally, Metz and colleagues' software calculates an AUC_z for the TOMM (see Fig. 2) of $.9814 \pm .0083$—virtually the same as the AUC of the best test represented in Fig. 3.

Tests That Satisfy the V&vG Criterion Can Be Made to Fail That Criterion

As the preceding paragraphs point out, an examiner has a choice of cutoffs when implementing or interpreting most malingering measures. Suppose an examiner thought that a good operating point for a malingering test was one at which FPR $\leq$ .15. These cutoffs are represented by rectangles on each ROC curve in Fig. 3. For the test represented by the lowest ROC curve, TPR = .565 when FPR = .15, and

[9]One can also calculate a "trapezoidal" or nonparametric AUC as determined by the area under a ROC graph in which (FPR, TPR) pairs are simply connected with line segments (for examples in the forensic mental health literature, see Douglas et al., 1999; Gardner et al., 1996; Rice & Harris, 1995). Concerning the mathematical evaluation of trapezoidal ROCs, see Mossman (1995).

LR+ = 3.77. For the middle ROC curve, TPR = .856 when FPR = .15, and $LR+ = 5.71$; for the upper ROC curve, TPR = .975 when FPR = .15, so that $LR+ = 6.50$. At these cutoffs, none of the tests satisfy the V&vG criterion, because a "positive" result raises the posttest probability of malingering from .30 to only .62 for the lower curve, .71 for the middle curve, and .74 for the upper curve. Thus, by simply adjusting the cutoff, an examiner can make the tests, all of which can satisfy the V&vG criterion, look like they fail to meet that standard.

Single Cutoffs Do Not Use All the Information in a Test Result

In assuming that malingering tests will be used with single cutoffs, the V&vG formulation risks losing some of the information in a test result. To see this, suppose that a diagnostician decides to designate a TOMM Trial 2 score below 45 as a "positive" test result. As was noted in Section 4.1, this cutoff choice sets TOMM's sensitivity in detecting malingering at .91, and its specificity at .95. Although these values represent excellent test accuracy, two things happen once this cutoff is selected. First, TOMM scores of 44 and 45 are treated very differently—one is deemed "positive" and the other "negative"—even though they reflect only the slightest difference in actual test performance. Second, TOMM scores of 44 and 34 are treated the same—both are deemed "positive"—despite the fact that a score of 34 represents much stronger evidence of malingering. Notice that such problems arise for a broad range of TOMM scores. Ideally, an interpretation of TOMM results should recognize that scores of 44 and 45 are similar, whereas scores of 34 and 44 give very different information about an evaluee.

Uncertainty in Estimating Positive Predictive Value

V&vG are not the first to suggest a Bayesian framework for evaluating malingering measures. Mossman and Hart (1996) suggested that Bayes's Theorem could help mental health professionals interpret data on malingering and make better presentations of evidence on malingering in court. However, Rogers and Salekin (1998) pointed out that imprecision in the estimated "base rate" of malingering (i.e., in $P(M+)$) and the estimated accuracy of malingering instruments (i.e., their sensitivity and specificity) would result in imprecision in $P(M+|T+)$, the posterior probability of malingering.[10] Although Rogers and Salekin believed that this imprecision necessitated abandonment of the Bayesian framework, Mossman (2000a, 2000b) and Mossman and Berger (2000) have described methods of producing intervals for $P(M+|T+)$, and concluded that the fears of Rogers and Salekin were unwarranted.

Nevertheless, the point Rogers and Salekin make about imprecision is important. To illustrate, suppose that a witness wished to assert that an evaluee was malingering on the basis of a test that showed that $P(M+|T+)$ equaled .90. Presumably, it would matter to a court to know whether the 95% credible interval for that estimate was, say, (.40–.98), or (.80–.97). If the former interval were correct, the court might feel that the data lacked adequate probative value, even in a case where the standard

[10]For earlier, general discussions of this point, see Monsour, Evans, and Kupper (1991), and Baron (1994).

of proof was only a "preponderance of the evidence." The latter interval, however, might satisfy a court's "clear and convincing" standard.

When proposing that a court accept test-based evidence of cognitive malingering, it would be desirable to describe that evidence while avoiding the flaws in V&vG's proposal. The following section does this. It describes a method for estimating posttest probability that takes into account a test's multiple outcomes and allows an evaluator to give a credible interval for the estimate.

5. DISTRIBUTIONS, CUTOFFS, AND CONFIDENCE INTERVALS

Equations 1 through 4 assume that the results of malingering measures can be dichotomized as either "positive" or "negative." The results of TOMM (shown in Fig. 1) suggest, however, that when investigators evaluate a malingering measure, they typically find that results of malingerers and honest responders form *distributions* rather than discrete categories. As the previous explains, it would be desirable to interpret test results in a way that recognizes and uses all the information in the distribution of test results. In the case of TOMM, we would like to interpret results such that gradually lower scores imply gradually stronger evidence of malingering.

To do this, we apply a method described in Mossman (2000a), and rewrite Eq. 1 as follows:

$$\mathrm{P}(M+|T_i) = \frac{\mathrm{P}(T_i|M+)\mathrm{P}(M+)}{\mathrm{P}(T_i|M+)\mathrm{P}(M+) + \mathrm{P}(T_i|M-)[1 - \mathrm{P}(M+)]} \tag{1a}$$

Equation 1(a) states that for a test T that has i possible outcomes, $\mathrm{P}(M+|T_i)$, the conditional probability of malingering given test outcome T_i is a function of

- $\mathrm{P}(M+)$, the pretest probability of malingering;
- $\mathrm{P}(T_i|M+)$, the probability of getting test result T_i if someone is malingering; and
- $\mathrm{P}(T_i|M-)$, the probability of getting test result T_i if someone is not malingering.

Applying to Eq. 1(a) algebraic procedures similar to those used to produce Eq. 4 yields

$$\mathrm{O}(M+|T_i) = LR_i \times \mathrm{O}(M+) \tag{4a}$$

Here, $\mathrm{O}(M+|T_i)$ is the *posterior odds* that an evaluee is malingering, given the test result T_i and pretest probability $\mathrm{P}(M+)$; $\mathrm{O}(M+)$ is the *prior odds*, and LR_i is the *likelihood ratio* associated with test result T_i.

From Eq. 4(a), it is obvious that imprecision in the estimates of LR_i and $\mathrm{O}(M+)$ will produce imprecision in the posterior odds. How might we calculate an interval for $\mathrm{O}(M+|T_i)$ that would express this imprecision quantitatively?

Mossman (2000a) has described a Monte Carlo method that will produce a useful interval for just this purpose. It involves (1) characterizing imprecision in the base rate of malingering, (2) characterizing imprecision in the likelihood ratio, and (3) combining these results mathematically to generate a distribution from which one can obtain an interval for $\mathrm{O}(M+|T_i)$.

Imprecision in the Base Rate of Malingering

Bayes's Theorem says that what one believes after receiving new information depends on what one believed before receiving that information. Thus, if one wishes to quantify one's confidence that an evaluee is malingering, given a particular test result, one has to quantify what one knew about the evaluee's probability of malingering before getting the test result. Here are three approaches to doing this.

(a) As V&vG note, several authors have offered estimates of malingering base rates that ranging from 7.5–15% (Trueblood & Schmidt, 1993) to 18–33% (Binder, 1993). People typically make systematic errors when attempting to estimate probabilities (Russo & Schoemaker, 1989). However, a clinician who thought these estimates applied to his evaluation setting might use a method of systematic self-interrogation suggested several years ago by Raiffa (1968). For example, the clinician might ask himself, "Given my knowledge and experience, what is the lowest possible rate of malingering in the population from which this evaluee is drawn? What is the highest possible rate? Can I give a range of rates such that I am 80 percent sure that the true rate of malingering falls within that range?" Following such a self-inquiry, the clinician might say that given his knowledge and experience, there is a 10% chance that the true base rate P($M+$) lies between .05 and .10, a 10% chance that $.30 \leq \mathrm{P}(M+) \leq .40$, and an 80% chance that $.10 \leq \mathrm{P}(M+) \leq .30$. The rectilinear distribution in Fig. 4 represents this clinician's beliefs.[11]

(b) A second approach would utilize a base-rate-determined "expert consensus." Rogers and Salekin (1998) gave an example of such findings, in which forensic clinicians were polled concerning their beliefs about the prevalence of malingering. The mean $\pm$ *SD* estimate was 0.1744 ± 0.1444. One can convert this estimate to a probability distribution by using the beta distribution, which provides a useful way of characterizing the a priori distribution of the parameter p of the binomial distribution, when p is the probability of a success on an individual trial (Iverson, 1984; Newcombe, 1998). One can describe the density of a $Beta(a, b)$ distribution as

$$\beta_{(a,b)}(p) = Cp^{a-1}(1-p)^{b-1} \tag{7}$$

for $0 < p < 1$, and zero otherwise (Freund & Williams, 1991); the constant C is used to normalize the total area under the distribution to equal 1. The relationship between the mean μ and standard deviation σ of a beta distribution and the parameters a and b is given by Iverson (1984):

$$a = \mu\left[\frac{\mu(1-\mu)}{\sigma^2} - 1\right] \qquad b = (1-\mu)\left[\frac{\mu(1-\mu)}{\sigma^2} - 1\right] \tag{8}$$

[11] This rectilinear distribution is imperfect: at either end, it suggests that probabilities are constant over a range, then drop suddenly to zero, implying that values just below or above the distribution are impossible. Such problems could be addressed by creating a rectilinear distribution with a more complicated shape, one representing the belief that all values between 0 and 1 were conceivable, but that some (e.g., those below .02 and above .6) are very unlikely.

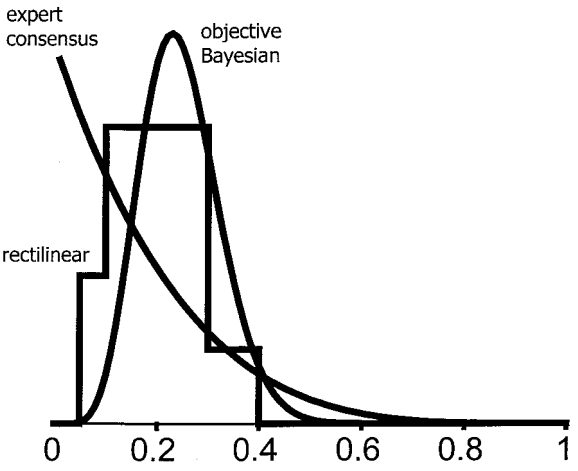

Fig. 4. Possible distributions for the pre-test probability of malingering.

Letting $\mu = 0.1744$ and $\sigma = 0.1444$, one can characterize the expert consensus, which also is shown in Fig. 4, as a beta distribution with $a = 1$ and $b = 5$.[12]

(c) A third approach utilizes actual available data, and characterizes it with "objective Bayesian" methods. Objective Bayesians use "noninformative" or default prior distributions that depend on the data model rather than on subjective beliefs. For a proportion p derived from an event that occurs x out of n times, Laplace (1812) recommended the "uniform" prior distribution, *Beta*(1, 1), and Jeffreys (1961) proposed the *Beta*(½, ½) prior distribution, now termed the "Jeffreys prior." Using the Jeffreys prior means that after observing x successes in n tries, we describe the posterior distribution of p as the *Beta*($x + ½$, $n - x + ½$) distribution. The virtues of using Jeffreys prior for p include the finding that it yields confidence sets with very good coverage properties (Brown, Cai, & DasGupta, 2001, 2002).

As an example, consider data reported by Gothard, Rogers, and Sewell (1995), who found that of 55 defendants who underwent evaluations of competence to stand trial, 25 were competent, 23 were incompetent, and 7 were malingering. Suppose one intends to evaluate malingering only among those defendants who appear incompetent (since the usual reason one tests for malingered incompetence is to sort defendants who are truly incompetent from those who are feigning illness). Then the relevant portion of Gothard and colleagues' sample does not include the 25 defendants who are deemed competent. Therefore, $\hat{p} = x/n = 7/30$; we represent this finding as the *Beta*(7½, 23½) distribution, also shown in Fig. 4.

[12]This characterization of the "expert consensus" is presented primarily for illustration purposes. Ten percent of the *Beta*(1, 5) distribution falls below .021. Yet it probably is incorrect to say either that 10% of experts think the rate of malingering among forensic evaluees falls below 2% or that experts think there is a 10% chance that the true rate of malingering among forensic evaluees falls below 2%.

Imprecision in the Likelihood Ratio

As explained by Mossman (2000a), the relationship among LR_i, the binormal ROC indices A and B, and the variance in LR_i, are described as follows. For a given score T_i on a malingering measure, LR_i is the ratio of the densities of the $M+$ and $M-$ distributions:

$$LR_i = B\,e^{-1/2\left[A^2+2ABz(FPR_i)+(B^2-1)z(FPR_i)^2\right]} \tag{9}$$

Taking the natural logarithm, Eq. 9 yields

$$\ln LR_i = \ln B - \frac{1}{2}\left[A^2 + 2ABz(FPR_i) + (B^2-1)z(FPR_i)^2\right] \tag{10}$$

It is a simple matter to use a polynomial to tightly fit test results (here, scores on TOMM) to values of $z(\mathrm{FPR}_i)$ (Somoza & Mossman, 1991b; Somoza, Soutullo-Esperon, & Mossman, 1989), so that for a particular test score, imprecision in $\ln LR_i$ is a function of imprecision in the estimates of A and B. By taking the partial derivatives of Eq. 10 with respect to A and B and using the Taylor series expansion (Elandt-Johnson & Johnson, 1980), one obtains this large-number estimate of the variance of $\ln LR_i$:

$$\begin{aligned}
\mathrm{var}(\ln LR_i) &\approx \left(\frac{\partial}{\partial A}\ln LR_i\right)^2 \mathrm{var}(A) + \left(\frac{\partial}{\partial B}\ln LR_i\right)^2 \mathrm{var}(B) \\
&\quad + 2\left(\frac{\partial}{\partial A}\ln LR_i\right)\left(\frac{\partial}{\partial B}\ln LR_i\right)\mathrm{cov}(A, B) \\
&= z(TPR_i)^2\,\mathrm{var}(A) + \left[\frac{1}{B} - z(FPR_i)z(TPR_i)\right]^2 \mathrm{var}(B) \\
&\quad - 2z(TPR_i)\left[\frac{1}{B} - z(FPR_i)z(TPR_i)\right]\mathrm{cov}(A, B)
\end{aligned} \tag{11}$$

In Eq. 11, $\mathrm{var}(A)$ and $\mathrm{var}(B)$ are the variances of A and B, and $\mathrm{cov}(A, B)$ is their covariance.[13] One can use the square route of this estimate for $\mathrm{var}(\ln LR_i)$ as the standard deviation of a distribution that characterizes imprecision in the estimate of LR_i. Thus, for example, the upper and lower limits of a $100(1-\alpha)\%$ interval[14] for LR_i would be estimated as

$$LR_{L,U} = e^{\left(\ln \mathrm{LR}_i \pm z_{(1-\alpha/2)}\sqrt{\mathrm{var}(\ln \mathrm{LR}_i)}\right)} \tag{12}$$

Imprecision in Posterior Odds and Posttest Probability

Having characterized the imprecision in the factors of the right-hand side of Eq. 4(a), one can implement a Monte Carlo procedure (see Mossman, 2000a;

[13] The variance–covariance matrix that is part of the output of Metz and colleagues' ROC analysis software contains $\mathrm{var}(A)$, $\mathrm{var}(B)$, and $\mathrm{cov}(A, B)$. For the TOMM data discussed earlier in this paper, $\mathrm{var}(A) = .2593$, $\mathrm{var}(B) = .0722$, and $\mathrm{cov}(A, B) = .1215$.

[14] For a 95% interval, one sets α equal to .05.

Table 2. Posterior Probabilities of Malingering for Three TOMM Scores (Trial 2), With 95% Intervals for Three Prior Probability Distributions

	TOMM scores		
	35	40	45
$z(\text{FPR}_i)$	−2.822	−2.244	−1.559
LR_i	46.7	8.23	1.14
Prior distribution			
Rectilinear	.919 (.723–.973)	.665 (.309–.876)	.214 (.058–.483)
Expert consensus	.878 (.197–.984)	.541 (.037–.915)	.147 (.005–.598)
Objective Bayesian	.934 (.817–.977)	.717 (.415–.891)	.260 (.094–.526)

Mossman & Berger, 2000) with a spreadsheet's random number generator to construct a distribution for $\text{O}(M+|T_i)$ and an estimate of the "equal-tailed" $100(1-\alpha)\%$ confidence set for $\text{O}(M+|T_i)$, which can be converted to a $100(1-\alpha)\%$ confidence set for $\text{P}(M+|T_i)$. That is, one uses a spreadsheet program to:

- draw, at random, one value from the distribution representing the prior probability of malingering, and convert this value to odds;
- draw, at random, one value from the distribution for $\ln LR_i$, and convert this value to LR_i
- combine these values in Eq. 4(a) to obtain a value of $\text{O}(M+|T_i)$;
- repeat this process N times (where N is a large number, say, 10,000) to generate N values of $\text{O}(M+|T_i)$;
- convert the N values of $\text{O}(M+|T_i)$ to values of $\text{P}(M+|T_i)$, and sort these values from lowest to highest;
- find the integer nearest to $N/2$, and choose the corresponding sorted value (i.e., the median of the N values) as the estimate of $\text{P}(M+|T_i)$;
- find the integers nearest to $N^{\alpha}/_2$ and $N(1-{}^{\alpha}/_2)$, and choose the $N^{\alpha}/_2$th and $N(1-{}^{\alpha}/_2)$th values as the lower and upper limits of the confidence set for $\text{P}(M+|T_i)$.

The results of this procedure appear in Table 2, which contains several results of interest. First, a Trial 2 TOMM score of 35 handily exceeds the V&vG criterion; because of its high likelihood ratio (46.7), all three central estimates look like they justify a high degree of confidence that an evaluee with this score is malingering. That confidence would be undermined, however, if the distribution labeled "expert consensus" in Fig. 4 were a valid representation of the prior probability of malingering. Using this prior distribution yields a 95% interval of (0.197–0.984). Even using the 90% interval (in this case, 0.317–0.978) only justifies the inference that the probability of malingering very likely is greater than 32%.

Second, a Trial 2 score of 40 increases the odds of malingering by a factor of 8—close to the V&vG criterion. Yet this score does not justify a high level of confidence that an evaluee is malingering, even though it lies well below most scores achieved by the "genuine" patients who made up the sample. If one examines the score distributions in Fig. 1, this finding is less of a surprise. There, one sees that relatively small fractions of both honest responders and malingerers scored in the

40–45 range, which suggests that what those scores say about an evaluee's status is ambiguous.

Finally, posttest ambiguity is at its greatest for evaluees who score 45. This score changes the odds of malingering very little, and imprecision in the initial base rate is amplified by imprecision in the indices of test accuracy.

CONCLUSION: COURTS, *DAUBERT*, AND EXPERTS' TEST DATA

One major function of a diagnostic medical test is to influence clinical decision making by influencing beliefs about the probability that a patient has a particular disorder. Test results from malingering measures have a similar role in a forensic context: the results influence fact-finders' decisions by influencing their opinions about the honesty of a litigant. Just as errors in medical diagnosis can have serious implications for treatment and outcome, erroneous judgments about malingering can undermine the accuracy and fairness of legal determinations about civil liability or criminal guilt (Rogers & Salekin, 1998). Professionals who present malingering data and fact-finders who assess those data therefore need to know how much a given test result should change their beliefs.

This paper has suggested that this problem is more complex mathematically than one might gather from reading V&vG's proposal for evaluating measures of cognitive malingering. The V&vG proposal does not recognize that malingering tests usually have more than two outcomes, and that because of this, what results from such measures entail changes gradually as one moves up or down the range of possible test outcomes. Moreover, V&vG ignore the ambiguity associated with Bayesian calculations of the likelihood of malingering.

An examination of data on TOMM's performance shows that errors in test-based judgments about malingering have at least two sources. First, malingering tests themselves, though very accurate, are likely to be imperfect, because the test results of honest responders overlap with those of test takers who are feigning or exaggerating. Even when a test result justifies a great increase in one's suspicion of malingering, some small likelihood may persist that the evaluee who produced the result was responding honestly. Second, ambiguity or imprecision is inherent in estimates of the pretest probability of malingering and the accuracy indices that characterize the malingering test. As a result, an evaluator's belief about the posttest probability of malingering is best characterized as an interval that can be calculated from (and that therefore incorporates) mathematical formulations of imprecision in base rates and accuracy indices.

The *Daubert* decision suggests that courts scrutinize "the known or potential rate of error" of scientific techniques that form the bases of proffered expert opinions. Although *Daubert* does not specify a minimally acceptable error rate, V&vG suggest that malingering measures might be evaluated against some such standard. This paper suggests, however, that the V&vG approach is unworkable. Malingering measures usually will have many possible scores, and therefore, many possible error rates. Therefore, the likelihood of error is a function not just of the test's properties, but of the score an evaluee actually achieves. Moreover, Bayes's Theorem requires that

test results be interpreted in light of available pretest knowledge. As applied to malingering measures, this means that evaluators must incorporate their numerical, but imprecise, estimates of pretest information—which may include a guess about the base rate of malingering and clinical data obtained from interviews or other sources—into determinations of what a test means. If, in making a judgment about a malingering test's "error rate," a court is concerned about the likelihood that an expert's conclusion is wrong, then Bayes's Theorem says the calculated error rate stems in part from the expert's pretesting information and beliefs about whether an evaluee was malingering, and from the ambiguity that results from efforts to express such information and beliefs in probabilistic terms.

A Final Comment

Although clinicians and investigators commonly use diagnostic procedures that were originally developed with the aid of complicated statistical operations, they rarely if ever have to perform those operations themselves. Even if clinicians embraced the conceptualization of test results that this paper presents, few of them would have the skills and inclination to carry out the mathematical process described in this paper. The author hopes this paper's description of test properties will inspire designers of tests to develop software that clinicians can use to create Bayesian interpretations of diagnostic findings.

REFERENCES

Baron, J. A. (1994). Uncertainty in Bayes. *Medical Decision Making*, *14*, 46–51.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–375.

Binder, L. (1993). Assessment of malingering after mild head trauma with the Portland Digit Recognition Test. *Journal of Clinical and Experimental Neuropsychology*, *15,* 170–182.

Brown, L., Cai, T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16,* 101–133.

Brown, L., Cai, T., & DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Annals of Statistics*, *30,* 160–201.

Clark v. State, 12 Ohio 483 (1843).

Coe v. Tennessee, 17 S.W.3d 193 (Tenn. 2000).

Connecticut Mutual Life Insurance Company v. Lathrop, 111 U.S. 612 (1884).

Cunnien, A. J. (1997). Psychiatric and medical syndromes associated with deception. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 23–46). New York: Guilford Press.

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

Dorfman, D. D., Berbaum K. S., Metz, C. E., Lenth, R. V., Hanley, J. A., & Abu-Dagga, H. (1997). Proper ROC analysis: The bigamma model. *Academic Radiology*, *4,* 138–149.

Douglas, K. S., Ogloff, J. R. P., Nicholls, T. L., & Grant, I. (1999). Assessing risk for violence among psychiatric patients: The HCR-20 violence risk assessment scheme and the psychopathy checklist: Screening version. *Journal of Consulting and Clinical Psychology*, *67,* 917–930

Downs v. Perstorp Components and ICI Americas, 126 F. Supp. 2d 1090 (E.D. Tenn. 1999).

Elandt-Johnson, R. C., & Johnson, N. L. (1980). *Survival models and data analysis*. New York: Wiley.

Frederick, R. I., & Crosby, R. D. (2000). Development and validation of the Validity Indicator Profile. *Law and Human Behavior*, *24,* 59–82.

Freund, J. E., & Williams, F. J. (1991). *Dictionary/outline of basic statistics*. Mineola, NY: Dover.

Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996). Clinical versus actuarial predictions of violence in patients with mental illnesses. *Journal of Consulting and Clinical Psychology*, *64,* 602–609.

Gothard, S., Rogers, R., & Sewell, K. W. (1995). Feigning incompetency to stand trial: An investigation of the Georgia Court Competency Test. *Law and Human Behavior*, *19,* 363–373.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1996). Comparison of multiple scoring methods for Rey's malingered amnesia measures. *Archives of Clinical Neuropsychology*, *11,* 283–293.

Guilmette, T. J., Hart, K. J., Giuliano, A. J., & Leininger, B. E. (1994). A comparison of the Fifteen Item Test and the Hiscock Forced-Choice Procedure in detecting simulated memory impairment. *Clinical Neuropsychologist*, *8,* 283–294.

Hanley, J. A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical Decision Making*, *8,* 197–203.

Hanley, J. A. (1996). The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine*, *15,* 1575–1586.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143,* 29–36.

Iverson, G. R. (1984). *Bayesian statistical inference.* Newbury Park, CA: Sage.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Clarendon Press.

Kumho Tire Co. v. Carmichael, 1526 U.S. 137 (1999).

Laplace, P. S. (1812). T*héorie analytique des probabilités*. Paris: V. Courcier.

McLean v. State, 16 Ala. 672 (1849).

Metz, C. E., & Pan, X. (1999). "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology*, *43,* 1–33.

Monsour, M. J., Evans, A. T., & Kupper, L. L. (1991). Confidence intervals for posttest probability. *Statistics in Medicine*, *10,* 443–456.

Mossman, D. (1994a). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62,* 783–792.

Mossman, D. (1994b). Further comments on portraying the accuracy of violence prediction. *Law and Human Behavior*, *18,* 587–593.

Mossman, D. (1995). Resampling techniques in the analysis of non-binormal ROCs. *Medical Decision Making*, *15,* 358–366.

Mossman, D. (2000a). The meaning of malingering data: Further applications of Bayes's Theorem. *Behavioral Sciences and the Law*, *18,* 761–779.

Mossman, D. (2000b). Interpreting clinical evidence of malingering: A Bayesian perspective. *Journal of the American Academy of Psychiatry and the Law*, *28,* 293–302.

Mossman, D., & Berger, J. O. (2000). Intervals for posttest probabilities: A comparison of 5 methods. *Medical Decision Making*, *21,* 498–507.

Mossman, D., & Hart, K. J. (1996). Presenting evidence of malingering to courts: Insights from decision theory. *Behavioral Sciences and the Law*, *14,* 271–291.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17,* 857–872.

Pankrantz, L., & Binder, L. M. (1997). Malingering on intellectual and neuropsychological measures. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 223–236). New York: Guilford Press.

Parry, J., & Drogan, E. Y. (2000). *Criminal law handbook on psychiatric and psychological evidence and testimony*. Washington, DC: American Bar Association.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under certainty*. New York: Random House.

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.

Resnick, P. J. (1997a). Malingered psychosis. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 47–67). New York: Guilford Press.

Resnick, P. J. (1997b). Malingering of posttraumatic disorders. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 130–152). New York: Guilford Press.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63,* 737–748.

Rogers, R. (Ed.). (1997a). Structured interviews and dissimulation. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 301–327). New York: Guilford Press.

Rogers, R. (Ed.). (1997b). Introduction. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 1–19). New York: Guilford Press.

Rogers, R., & Salekin, R. T. (1998). Beguiled by Bayes: A reanalysis of Mossman and Hart's estimates of malingering. *Behavioral Sciences and the Law*, *16,* 147–153.

Russo, J. E., & Schoemaker, P. J. H. (1989).*Decision traps: the ten barriers to brilliant decision-making and how to overcome them*. New York: Simon and Schuster.

Schretlen, D., Brandt, J., Krafft, L., & van Gorp, W. G. (1991). Some caveats in using the Rey 15-Item Memory Test to detect malingered amnesia. *Psychological Assessment*, *3,* 667–672.
Somoza, E., & Mossman, D. (1991a). ROC curves and the binormal assumption. *Journal of Neuropsychiatry and Clinical Neurosciences*, *3,* 436–439.
Somoza, E., & Mossman, D. (1991b). "Biological markers" and psychiatric diagnosis: Risk–benefit balancing using ROC analysis. *Biological Psychiatry*, *27,* 811–826.
Somoza, E., Soutullo-Esperon, L., & Mossman, D. (1989). Evaluation and optimization of diagnostic tests using ROC analysis and information theory. *International Journal of Biomedical Computing*, *24,* 153–189.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000, October). Better decisions through science. *Scientific American*, *283,* 82–87.
Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic tests: Methods from signal detection theory*. Orlando, FL: Academic Press.
Tombaugh, T. N. (1996). *TOMM: Test of memory malingering*. North Tonawanda, NY: Multi-Health Systems, Inc.
Tombaugh, T. N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, *9,* 260–268.
Trueblood, W., & Schmidt, M. (1993). Malingering and other validity considerations in the neuropsychological evaluation of mild head injury. *Journal of Clinical and Experimental Neuropsychology*, *15,* 578–590.
United States v. Gigante, 982 F. Supp. 140 (E.D. N.Y. 1997a).
United States v. Gigante, 996 F. Supp. 194 (E.D. N.Y. 1997b).
Vallabhajosula, B., & van Gorp, W. G. (2001). Post-*Daubert* admissibility of scientific evidence on malingering of cognitive deficits. *Journal of the American Academy of Psychiatry and the Law*, *29,* 207–215.
Villalba v. Consolidated Freightways Corp., 2000 U.S. Dist. LEXIS 11773 (N.D. Ill., August 10, 2000).