Behavioral Sciences and the Law
Behav. Sci. Law **18**: 693–718 (2000)

Mixed Group Validation: A Method to Address the Limitations of Criterion Group Validation in Research on Malingering Detection

Richard I. Frederick, Ph.D.*

Mixed group validation (MGV) is offered as an alternative to criterion group validation (CGV) to estimate the true positive and false positive rates of tests and other diagnostic signs. CGV requires perfect confidence about each research participant's status with respect to the presence or absence of pathology. MGV determines diagnostic efficiencies based on group data; knowing an individual's status with respect to pathology is not required. MGV can use relatively weak indicators to validate better diagnostic signs, whereas CGV requires perfect diagnostic signs to avoid error in computing true positive and false positive rates. The process of MGV is explained, and a computer simulation demonstrates the soundness of the procedure. MGV of the Rey 15-Item Memory Test (Rey, 1958) for 723 pre-trial criminal defendants resulted in higher estimates of true positive rates and lower estimates of false positive rates as compared with prior research conducted with CGV. The author demonstrates how MGV addresses all the criticisms Rogers (1997b) outlined for differential prevalence designs in malingering detection research.

INTRODUCTION

A clinician administers a test. A ''positive'' test score indicates the presence of pathology (or some other condition of interest). The probability that an individual with pathology will earn a positive test score is the ''true positive rate'' (TPR) of the

*Correspondence to: Richard I. Frederick, Ph. D., Department of Psychology, U.S. Medical Center for Federal Prisoners, Springfield, Missouri, 65807, USA. E-mail: rfrederi@ipa.net
I thank Robyn Dawes, who reviewed an earlier version of this manuscript, and Paul Meehl, who graciously provided assistance in researching this topic. I also thank forensic examiners at the U.S. Medical Center for Federal Prisoners for their assistance: David Mrad, Christina Pietz, Robert Denney, James Wolfson, and Richart DeMier. I appreciate Douglas Mossman's assistance in explaining certain methodology in receiver operating characteristic curve analysis. This paper does not necessarily represent the views or official policies of the U.S. Department of Justice or the Federal Bureau of Prisons.

test score. The probability that an individual without pathology will earn a positive score is the ''false positive rate'' (FPR) of the test score. The rate of observing positive scores in a sample of individuals with and without pathology is given as:

$$\begin{aligned}\text{proportion positive scores} &= (\text{TPR} \times \text{base rate pathology}) \\ &\quad +(\text{FPR} \times \text{base rate no pathology}), \text{ or} \\ \text{S+} &= \text{TPR(P+)} + \text{FPR(P-)}.^{1} \qquad (1)\end{aligned}$$

For example, if a sample has a rate of ''pathology''=.80, then the rate of ''no pathology''=.20. If a test with TPR=.90 and FPR=.15 is administered, the rate of positive scores (S+) will be .75; S+=.90(.80)+.15(.20)=.75.

Researchers try to find cutoff scores that maximize true positive scores and minimize false positive scores. Researchers commonly design validation studies for test cutoff scores by creating two groups, each with a different rate of pathology. Generally, according to inclusion and exclusion criteria developed for this purpose, one group is formed to contain no members with pathology (negative criterion group; P+=0), and one is established so that all members exhibit pathology (positive criterion group; P+=1). Testing with the new test is conducted. The proportion of positive scores in the positive criterion group estimates the TPR of the new test, and the proportion of positive scores in the negative criterion group estimates the FPR (see Table 1). This process is commonly called criterion group validation (CGV).

Mixed group validation (MGV) (Dawes & Meehl, 1966) is a process of using ''mixed groups'' to estimate the TPR and FPR of test scores. ''Mixed'' means that validation groups contain a mixture of individuals both with and without pathology, as opposed to criterion groups, which are assumed to contain only one type of individual. To conduct MGV, validation groups must also exhibit different rates of pathology (P+), otherwise there would be no basis to expect differences in rates

Table 1. Computation of diagnostic efficiencies in criterion group validation

	Clinical condition	
Test score	Pathology present	Pathology absent
Positive	A True positives	B False positives
Negative	C False negatives	D True negatives

True positive rate=probability positive test score when pathology present
TPR=A/(A+C)

False positive rate=probability positive test score when pathology absent
FPR=B/(B+D)

Positive predictive power=probability positive test score represents pathology
PPP=A/(A+B)

Negative predictive power=probability negative test score represents absence of pathology
NPP=D/(C+D).

[1] (P+)+(P−)=1: the proportion of individuals with and without pathology accounts for all observations within a sample. (S+)+(S−)=1. TPR and FPR are sometimes referred to as ''sensitivity'' and ''nonspecificity,'' respectively. Nonspecificity is equal to (1−specificity).

of positive test scores. Unlike CGV, the different values of P+ do not have to be 0 and 1. Although knowing the rate of pathology in a mixed group is required, it is not necessary to know which of the *individual* participants within any validation group manifests the pathology.

Consider two mixed groups. Rates of pathology within each group are labeled P_1+ and P_2+. The test of interest is administered. The proportion of group members in each validation group who earn a positive score is given by S_1+ and S_2+. Equation (1) is adapted for each group:

$$S_1+ = TPR(P_1+) + FPR(P_1-)$$

$$S_2+ = TPR(P_2+) + FPR(P_2-)$$

TPR and FPR are posited as stable characteristics of the test within both groups. Assuming that TPR exceeds FPR, if group 1 has a higher rate of pathology than group 2 (i.e., $P_1+ < P_2+$), then there should be a higher rate of positive test scores in group 1 than in group 2 (i.e., $S_1+ < S_2+$). Changes in the rate of pathology result in predictable changes in the rate of positive test scores. To elucidate the relationship among P+, S+, TPR, and FPR, equation (1) can be re-written as

$$S+ = TPR(P+) + FPR(1 - P+)$$

which can be re-written as

$$S+ = FPR + (TPR - FPR)P+ \quad (2)$$

This is a linear equation in the form of

$$y = a + bx.$$

Equation (2) indicates that when the *proportion of pathology* within a mixed group (P+) is plotted as the x-value against the observed *proportion of positive test scores* within the mixed group (S+, the y-value), then the FPR is equal to the y-intercept at $x=0$ (i.e., "a"). The slope of the line of best fit (i.e., "b") between coordinates of (P+ and S+) for each validation group is equal to TPR −FPR; consequently, the TPR is equal to the y-intercept at a line drawn for $x=1$. Figure 1, panel A, represents this relationship.

In Figure 1, the values of y at $x=0$ and $x=1$ represent CGV. That is, FPR is the rate of positive test scores when pathology is absent (i.e., $x=0$) and TPR is the rate of positive test scores when only pathology is present (i.e., $x=1$). A line drawn between these points represents MGV, representing the spectrum of potential mixed groups of individuals with and without pathology. Along this line, it is possible to estimate the proportion of positive scores in groups, if the rates of pathology are known (Figure 1, panel B). Conversely, along this line, the rates of pathology can be estimated from the proportions of positive scores in a group (Figure 1, panel C). Furthermore, the proportions of pathology and positive scores observed in any two groups (e.g., P_1+, S_1+ and P_2+, S_2+; $P_1+ \neq P_2+$) can be

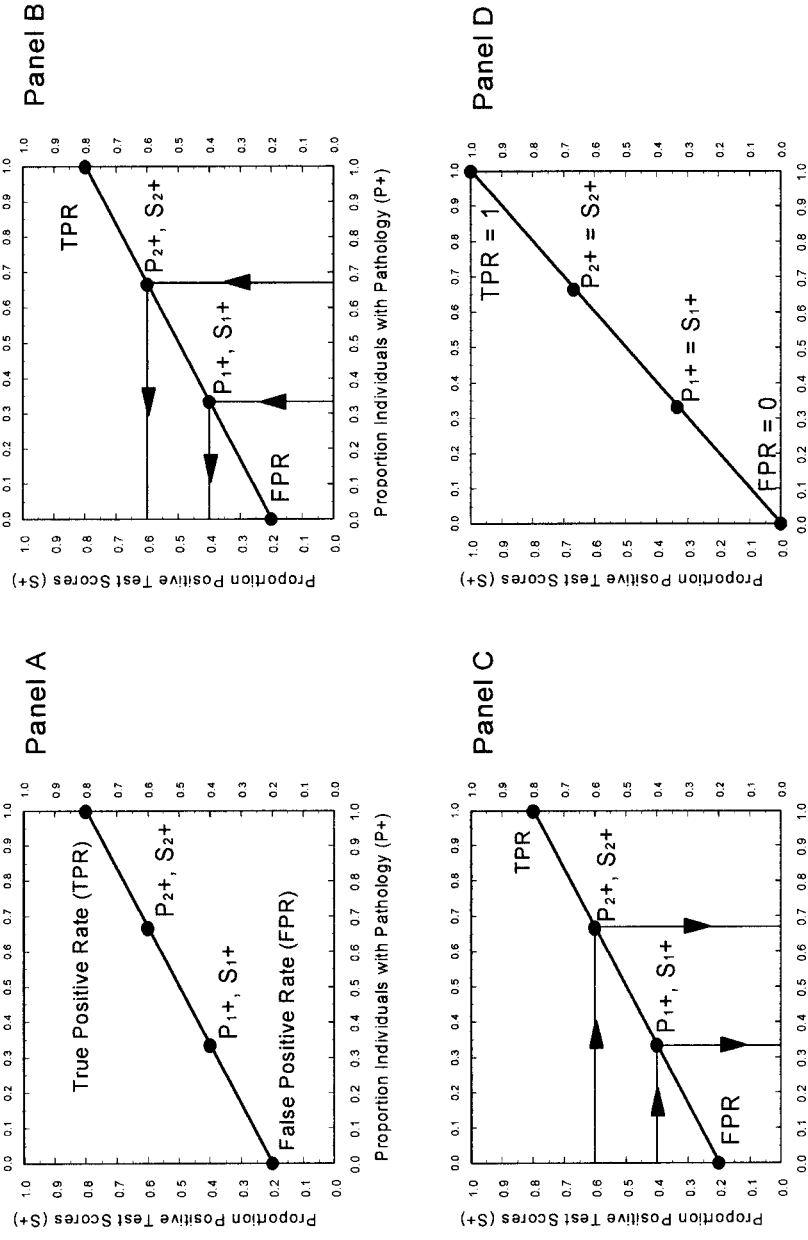

Figure 1. Mixed group validation (MGV) is based on a linear relationship among the rates of pathology (P+), the rates of positive test scores (S+), and the true positive rate (TPR) and false positive rate (FPR) of the test under examination. In panel A, a line is drawn through the points representing the rates of pathology (P_1+ and P_2+) and the corresponding rates of positive tests scores (S_1+ and S_2+). The line crosses the *y*-axis at $x=0$ and at a line drawn for $x=1$. The *y*-values at these crossing points represent the FPR and TPR of the test. If the FPR and TPR of the test are already known, one can predict S_1+ and S_2+ if one also knows P_1+ and P_2+ (panel B). Conversely, if one knows the FPR, TPR, S_1+, and S_2+, then one can estimate P_1+ and P_2+ (panel C). When $P+=S+$ in two samples and $P_1+\neq P_2+$, then one has a perfect test; FPR = 0 and TPR = 1 (panel D).

plotted as two points. A line drawn between the two points that extends through the axes at $x=0$ and $x=1$ estimates TPR and FPR. An interesting characteristic of these relationships is that when, within each of two mixed groups with different rates of pathology, the proportion of positive test scores equals the rate of pathology, the line will cross $x=0$ at $y=0$ and will cross $x=1$ at $y=1$, representing FPR=0 and TPR=1, describing a perfect test (Figure 1, panel D).

Goodman (1953, 1959) first described this method of investigation to cope with social science questions for which group data were readily available, but individual data were not (his example was the nature of the relationship between illiteracy and racial groups). Goodman proposed the method as an economical way to discover the efficiencies of predictor variables. MGV was independently proposed by Dawes and Meehl (1966) to address limitations of CGV. Their goal was to introduce a method to investigate constructs for which is was impossible to establish criterion groups, immediately (e.g., future suicidality) or definitively (e.g., schizotaxia, Meehl, 1995). Dawes and Meehl offered a algebraic solution for two groups, beginning with these equations:

$$S_1+ = TPR(P_1+) + FPR(P_1-)$$

$$S_2+ = TPR(P_2+) + FPR(P_2-)$$

By means of simultaneously solving these two equations, the TPR and FPR of the test score were computed:

$$TPR = \frac{[S_2+ \times P_1-] - [S_1+ \times P_2-]}{(P_2+) - (P_1+)} \tag{3}$$

$$FPR = \frac{[S_1+ \times P_2+] - [S_2+ \times P_1+]}{(P_2+) - (P_1+)} \tag{4}$$

Computer Simulation of Mixed Group Validation

The process of MGV is illustrated by means of a computer simulation. The simulation was designed to demonstrate that one can estimate the TPR and FPR for a test when one knows the rate of pathology and the proportion of positive test scores within two mixed groups, but one does not know the status of individuals within the groups with respect to the presence or absence of pathology. To begin, a test with TPR=.70 and FPR=.05 was hypothesized. Data sets were prepared to represent the distribution of test scores for two pure criterion groups of 1000 persons with and without pathology. Test scores were represented by ones (positive test scores, consistent with pathology) or zeros (negative test scores, consistent with no pathology). Test scores within ''pure group 1,'' the data set representing ''pathology present,'' comprised 700 ones and 300 zeros (i.e., TPR=.70). Test scores within ''pure group 2,'' the data set representing ''pathology absent,'' included 50 ones and 950 zeros (i.e., FPR=.05).

Data sets representing mixed groups of individuals with and without pathology were formed by randomly sampling test scores from each of the pure group data sets until 1000 test scores were drawn. Mixed group data sets always retained knowledge of the proportion of test scores drawn from each pure group, but never retained knowledge of which ones or zeros had come from which pure group. Nine *types* of mixed group were formed with ratios of pathology to non-pathology at 9:1, 8:2, and so on, through 1:9. For example, to form a mixed group with a 6:4 ratio of pathology to non-pathology, 600 of the observations were randomly drawn (without replacement) from the pure group 1 data set, and 400 of the observations were drawn from the pure group 2 data set. This process was repeated 100 times for each type of mixed group, for a total of 900 mixed groups. Once mixed group data sets were created, there was no way of identifying which zeros and ones had come from which pure group data set; however, the ratio of scores from pure group 1 to scores from pure group 2 was always known (e.g., $P+$ for any 6:4 group was equal to 0.6).

Simulation of MGV began once all 900 mixed groups had been created. One sample of the 9:1 mixed groups was compared to one sample of the 8:2 mixed groups. Comparisons computed the six values (P_1+, P_1-, S_1+, P_2+, P_2-, and S_2+) needed to calculate the TPR and FPR by comparing two mixed groups, according to the equations (3) and (4). Four of these values were easily observed based on the types of sample compared (i.e., in this first comparison, $P_1+=.9$, $P_1-=.1$, $P_2+=.8$, $P_2-=.2$). To derive S_1+ and S_2+ for the mixed groups used within a comparison, the number of ones within each mixed group was summed and the total was divided by 1000 (the total number of test scores within each mixed group). After the six values had been derived, TPR and FPR were calculated for the comparison using equations (3) and (4). This was followed by a comparison of another sample of a 9:1 mixed group with another sample of a 8:2 mixed group, deriving another estimate of TPR and FPR. All 100 samples of 9:1 mixed groups were compared to different samples of 8:2 mixed groups, deriving 100 estimates of TPR and FPR for the hypothetical test. Then the 9:1 mixed groups were compared with the 7:3 mixed groups, and so forth, across the 36 possible combinations of comparison, until the 2:8 mixed groups were compared with the 1:9 mixed groups. All 36 possible combinations included 100 comparisons, yielding 3600 estimates of TPR and FPR.

Results

Mean TPR was .70 ($n=3600$, SD$=.04$; range$=-.24$ to 1.44; one estimation of TPR < 0 and two estimations of TPR > 1); mean FPR was .05 ($n=3600$, SD$=.05$; range$=-.27$ to .28; 175 estimations of FPR < 0). Table 2 reports the mean proportion of positive scores ($S+$) observed for each type of mixed group. Table 3 reports the mean TPR and FPR estimations for each type of comparison. Figures 2 presents histograms of the 3600 estimations of TPR and FPR.

Figure 3 shows the prevalence of pathology within each type of mixed group, plotted against the average positive test sign rates for each type of mixed group (from Table 2). The y-intercept at $x=0$, which represents the FPR, is 0.05. The value of the y-intercept at $x=1$, representing the TPR, is 0.70. These are the same values as established *a priori* for the hypothetical test.

Table 2. Mean proportion of positive test scores observed in each type of mixed group in computer simulation

Ratio of pure groups within mixed group (P+)[a]	Predicted proportion positive scores[b]	Observed mean proportion positive scores (S+)[c]	SD
9:1	.635	.636	.005
8:2	.570	.570	.006
7:3	.505	.505	.008
6:4	.440	.440	.008
5:5	.375	.376	.007
4:6	.310	.309	.009
3:7	.245	.244	.009
2:8	.180	.178	.013
1:9	.115	.115	.005

Note: [a]Values represent the proportion of members with pathology to member without pathology. A 9:1 mixed group contained 90% members with pathology, 10% without pathology.
[b]The predicted proportion of positive scores in any mixed group is derived by adding the product of the true positive rate (TPR) and proportion of members with pathology to the product of the false positive rate (FPR) and proportion of members without pathology (see equation (1), in text. Because the hypothesized TPR = .70 and hypothesized FPR = .05, for a 9:1 mixed group, the predicted proportion of positive scores is (.9)(.7) + (.1)(.05) = .63 + .005 = .635.
[c]Means represent 100 samples of each type of mixed group.

Table 3. Mean true positive rate (TPR) and false positive rate (FPR) for each type of mixed group comparison in computer simulation

Type of comparison	TPR	FPR	Type of comparison	TPR	FPR
9:1 to 8:2	.702	.040	7:3 to 3:7	.700	.049
9:1 to 7:3	.701	.046	7:3 to 2:8	.701	.048
9:1 to 6:4	.701	.049	7:3 to 1:9	.700	.050
9:1 to 5:5	.701	.050	6:4 to 5:5	.699	.052
9:1 to 4:6	.701	.048	6:4 to 4:6	.702	.048
9:1 to 3:7	.701	.049	6:4 to 3:7	.701	.049
9:1 to 2:8	.701	.048	6:4 to 2:8	.702	.048
9:1 to 1:9	.701	.050	6:4 to 1:9	.700	.050
8:2 to 7:3	.699	.051	5:5 to 4:6	.706	.045
8:2 to 6:4	.699	.052	5:5 to 3:7	.703	.048
8:2 to 5:5	.699	.052	5:5 to 2:8	.704	.047
8:2 to 4:6	.700	.049	5:5 to 1:9	.701	.050
8:2 to 3:7	.700	.049	4:6 to 3:7	.699	.050
8:2 to 2:8	.700	.048	4:6 to 2:8	.702	.048
8:2 to 1:9	.700	.048	4:6 to 1:9	.698	.050
7:3 to 6:4	.698	.053	3:7 to 2:8	.707	.046
7:3 to 5:5	.699	.052	3:7 to 1:9	.698	.050
7:3 to 4:6	.700	.049	2:8 to 1:9	.686	.052

Note: Mean TPRs and FPRs are based on 100 comparisons between the two types of mixed group using the Dawes Meehl (1966) equations (equations (3) and (4), see text). For the comparison of 9:1 (group 1) to 5 : 5 (group 2), the first mixed group had 90% members with pathology and the second group had 50% members with pathology.

Discussion

Knowledge of the source of individual test scores was not required to obtain reliable estimates of the TPR and FPR of the hypothesized test. Ninety-five percent of estimates of the TPR were within 0.09 of its true value (.61 to .79) and 95% of the FPR estimates were within 0.10 of its true value (−.05 to .15).

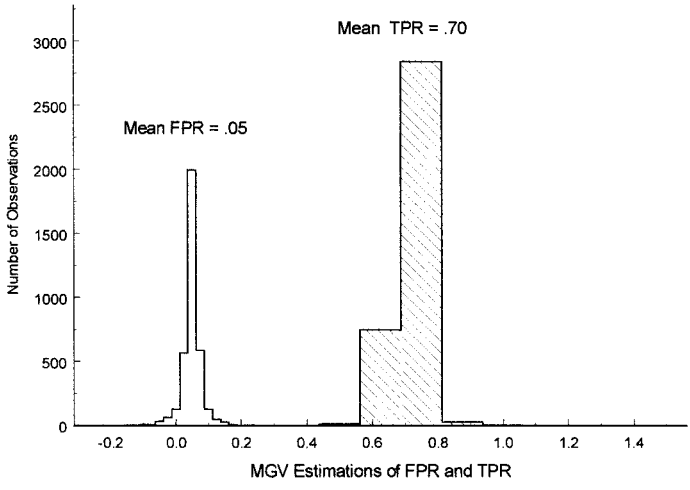

Figure 2. Histograms for 3600 estimations of FPR and TPR in the computer simulation. FPR was set to 0.05 and TPR was set to 0.70 prior to estimation. Some instance of impossible probability values occurred (less than zero or greater than one). The average derived value of FPR was 0.05; the average derived value of TPR was 0.70.

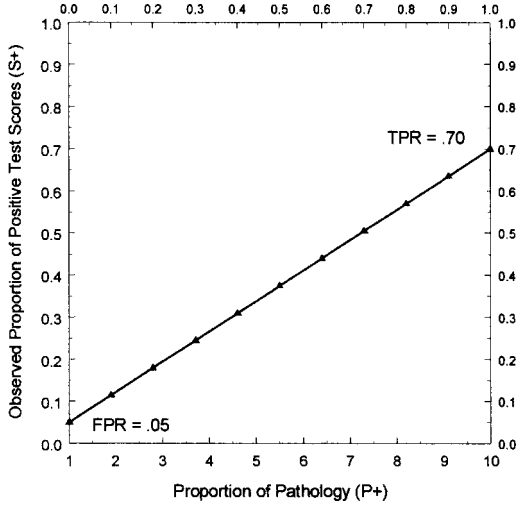

Figure 3. Mean observed proportions of positive test scores (S+) were plotted at each corresponding rate of pathology (P+) for the mixed groups used in the computer simulation. When a line is drawn through the values, the line crosses the y-axis at 0.05 (at $x=0$, $y=0.05$) and crosses a line drawn at $x=1$ (where $y=0.70$). The observed y values are the respective values of the FPR and TPR for the hypothesized test.

Some impossible probability values resulted; there were 175 negatively valued estimations for FPR. This is not surprising, given the near-zero value of FPR. Such outcomes were of great concern to Alf and Abrahams (1967) and Linn (1967), but the occurrence is no more condemning of MGV than the occurrence of an F ratio < 1 that negates the utility of analysis of variance (Rorer & Dawes, 1982). When estimates of TPR and FPR are outside the bounds of 0 and 1, Goodman (1959) recommended checking underlying assumptions about the constancy of TPR and FPR. Potential violations of the assumption include situations in which the probability of a positive test score is a more a function of some condition other than the

presence of pathology. When the assumptions of constancy are warranted, Goodman recommended truncating impossible values to 0 or 1.

MGV requires knowledge of the *rate* of pathology within mixed groups. CGV requires *individual* decisions about the presence of pathology for each participant. Consequently, CGV is typically a time-intensive and expensive process. MGV has been available for over 40 years. Nevertheless, MGV has rarely been incorporated into research designs (examples include Cobb, Hunt, & Harburg, 1969; Knowles & Schroeder, 1990) despite the availability of numerous source of prevalence data regarding psycholegal issues. Several potential objections may have contributed to this cool reception.

(1) "*An essential underpinning of MGV, the constancy of TPR and FPR, is unfounded.*" Dawes (1967) addressed the issue of constancy of TPR and FPR with respect to a test for schizophrenia:

> What the assumption states is that the *group* to which a [person with schizophrenia] belongs should not affect the probability of having a certain schizophrenic symptom . . . this assumption is also implicit in the orthodox manner of assessing statistical contingency whenever we speak of *the* contingency between symptom and schizophrenia without specifying certain subgroups of [persons with or without schizophrenia] we have in mind (p. 404, emphasis in the original).

Dawes recognized that variation related to sampling error in estimating TPR and FPR for a test would be observed across different samples (see also Baldessarini, Finklestein, & Arana, 1983; Goodman, 1959). The assumption of constancy of TPR and FPR is an implicit, but essential, underpinning of CGV; the assumption is merely overtly stated for MGV. Constancy of TPR and FPR is assumed by test consumers who administer tests to individuals who are outside the original validation groups (Baldessarini et al., 1983; Dawes, 1967; Elwood, 1993; Meehl & Rosen, 1955). Neither validation method should be used when the assumption of constancy TPR and FPR is untenable.

In clinical practice, test users are most interested in positive predictive power (PPP) and negative predictive power (NPP). PPP is the probability that a positive test score correctly predicts pathology; NPP is the probability that a negative test score correctly predicts the absence of pathology. These values *do not remain constant* and vary substantially according to the prevalence of pathology (Baldessarini et al., 1983; Dawes, 1962; Elwood, 1993; Meehl & Rosen, 1955; Mossman & Somoza, 1991). As the prevalence of pathology approaches zero, the PPP approaches zero, even for highly sensitive tests: If no pathology is present, all positive scores will be in error.[2]

[2]Rogers, Sewell, Cruise, Wang, & Ustad (1998) reported that TPR and FPR "are highly contingent on base rates" (p. 403). Their error is not unique. For example, Butcher, Graham, and Ben-Porath (1995) discussed the process of establishing the validity of cutoff scores on the of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). In an apparently hypothetical example regarding the MacAndrew Alcoholism Scale—Revised (MAC-R), Butcher et al. (1995) reported a TPR of 0.90 and FPR of 0.10 for MAC-R when the base rate of alcohol abusers was 0.50, but conjectured that the FPR increased to 0.90 when the base rate dropped to 0.20 (Figure 2, p. 327). Based on this egregious assumption, Butcher et al. (1995) claimed that the TPR and FPR were not primary in test development to establish cutoff scores. Instead, they insisted that PPP and NPP should be "the focus of MMPI-2 researchers" (p. 327), a conclusion fortunately disputed by Nicholson, Mouton, Bagby, Buis, Peterson, & Buigas (1997).

(2) "*It is necessary to know which individuals in a mixed group do or do not manifest pathology before the TPR and FPR can be estimated.*" Hopefully, this concern, which is a vestige of attempts to purify criterion groups in CGV, was allayed by the computer simulation. In CGV, researchers work to approach rates of pathology within criterion groups of 0% and 100%. In this vein, they are right to carefully screen all participants. Nevertheless, they ultimately have little concern about which individuals in the groups earn positive or negative test scores when they compute TPR and FPR. That is, TPR and FPR are computed based on *cell* counts, not on *individual* counts.

Because there exists a *linear* relationship among TPR, FPR, P+, and S+, and because TPR and FPR are constants, changes in S+ must reflect proportional changes in P+. If changes in the rate of positive test scores are not predictably proportional to changes in the base rate of pathology, then the test has no utility. Because these proportional changes can be predicted solely from group data, it is not necessary to know individual contributions to the proportion of observed positive scores.

(3) "*Estimates of pathology within mixed groups are subject to error.*" Estimations of pathology within validation groups are subject to error in both CGV and MGV. In CGV, researchers attempt to establish criterion groups comprising only one type of member. Failure to establish criterion groups that are completely pure constitutes an overestimation of P+ for the positive criterion group (which results in an underestimation of the TPR) and an underestimation of P+ for the negative criterion group (which yields an overestimation of the FPR). In essence, CGV estimation of TPR and FPR for a new test cannot exceed the TPR or FPR of the inclusion and exclusion rules that establish criterion groups, except by error. Consequently, only when *perfect* criterion group inclusion and exclusion criteria are available will CGV avoid significant error in estimation of the proportion of pathology within each group. Criterion group contamination is a common problem, particularly in early attempts at validating test scores or in investigating new constructs, because the criteria used to assign potential research participants to groups are typically weakly valid.

As an example, consider inclusion and exclusion criteria with an observed TPR of 0.8 and FPR of 0.2. These criteria are used to establish criterion groups in order to validate a new test, which happens to be *perfect* (i.e., TPR=1.0 and FPR=0.0). Table 4 shows what will happen by means of CGV. For the perfect test, the TPR is underestimated at 0.8 and the FPR is overestimated at 0.2. The perfect test is reported to be imperfect; progress is impeded.

MGV, on the other hand, effectively uses even weakly valid pathology indicators to validate better indicators. Figure 4 shows the process of MGV for the same example. The FPR (0.8) and TPR (0.2) of the inclusion and exclusion criteria are plotted at $x=0$ and $x=1$, respectively. A line is drawn between them. The rate of positive signs for the inclusion and exclusion criteria, S_1+, is observed to be 0.40. By finding its coordinate value at the line drawn between $y=0.2$ and $y=0.8$, P_1+ is estimated to be .33. For group 2, $S_2+=.60$; consequently, P_2+ is estimated to be .67 by observing its coordinate value at the line. The perfect test is administered. Because the test is perfect, the sign rate of the new test for group 1 is .33; the sign rate for group 2 is .67. Rates of positive scores perfectly match the base rates within groups 1 and 2. A line

Table 4. Validation of a perfect test using criterion groups validation with imperfect inclusion and exclusion criteria

	Assumption of pathology based on imperfect criteria		
Perfect test sign	Assumed present	Assumed absent	Total
Positive	80	40	120
Negative	20	160	180
Total	100	200	300

Computed true positive rate (TPR) = 80/100 = .80. Computed false positive rate (FPR) = 40/200 = .20

	Actual pathology within criterion groups				
	Assumed present		Assumed absent		
Perfect test sign	Actual present	Actual absent	Actual present	Actual absent	Total
Positive	80	0	40	0	120
Negative	0	20	0	160	180

Note: The upper table demonstrates traditional criterion groups validation when criterion groups have been formed with imperfect inclusion and exclusion criteria (TPR = .80, FPR = .20). Twenty percent of the members of the criterion groups are placed in the wrong categories. Consequently, a perfect test is also determined to have TPR = .80 and FPR = .20. The lower table reveals the true status of members, demonstrating that the perfect test actually categorized each individual correctly despite the researcher's inability to observe the process.

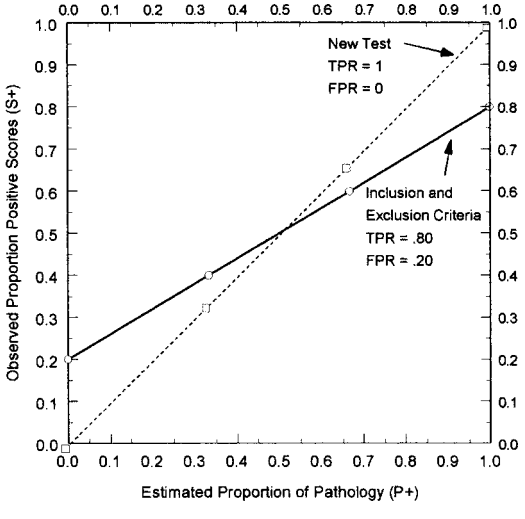

Figure 4. An imperfect test (FPR = .20; TPR = .80) is used to validate a perfect test. The FPR and TPR are plotted and a line is drawn through them. The proportion of positive scores on the imperfect test for two samples (0.40 and 0.60) estimate the rate of pathology (P+) in the groups as 0.33 and 0.67, respectively. The two samples generate positive scores for the perfect test at the rates of 0.33 and 0.67. The values for (P+, S+) are plotted at (.33, .33) and (.67, .67). A dashed line drawn through these points estimates FPR = 0.0 and TPR = 1.0, the values of a perfect test.

drawn between these new points crosses $x=0$ at $y=0$ (thus FPR = 0) and crosses $x=1$ at $y=1$ (thus TPR = 1). The sort of error that weakens CGV presents no limitation for MGV.

Applying Mixed Group Validation for Malingering Research

It is difficult to establish pure criterion groups for research concerning the diagnostic efficiencies of tests that purport to detect malingered cognitive impairment. Rogers (1997b) has promoted the use of ''known groups'' (clinical criterion groups) in malingering research because of the superior generalizability of such designs over the use of ''simulators'' (analog criterion groups in which individuals play roles). Within simulation designs, researchers can only hope that participants perform as instructed; often it turns out that many do not (Arnett, Hammeke, & Schwartz, 1995; Frederick, Sarfaty, Johnston, & Powel, 1994; Goebel, 1983).

Nevertheless, it is only with great optimism that one may speak of a ''known groups'' design, given that the nature of malingering research typically precludes one from ''knowing'' the true status of clinical participants (Greiffenstein, Baker, & Gola, 1994, 1996; Greiffenstein, Gola, & Baker, 1995; Viglione, Fals-Stewart, & Moxham, 1995). Without perfect criteria, one can never know a participant's presentation is genuine.

Additionally, the rubric of ''known groups'' is somewhat misleading, for it is unclear whether Rogers (1997b) intended ''known groups'' to reflect perfect confidence in group memberships. Rogers et al. (1998), for example, reported a ''known-groups'' comparison in which the criterion for membership in the ''known'' positive group had an estimated TPR of .98 and the criterion for membership in the ''known'' negative group had an estimated FPR of .05. Consequently, ''known'' groups formed by these criteria will be impure; a small percentage of individuals will be placed in the wrong criterion group. This process actually constitutes the ''differential prevalence'' design, one in which the status of individual members is uncertain and there exists a different rate of pathology within the two groups. Without recognizing this distinction, Rogers (1997b) has excoriated the differential prevalence design:

> We learn very little from differential prevalence designs. By design, we not know *who* is dissimulating in each group. Logically, we do not know *how many* are dissimulating in each group. Even when groups yield predicted differences, we do not know *what meaning* should be assigned to deviant or atypical scores. For all we know, every ''deviant'' or ''atypical'' score could be indicative of honest responding. We also do not know *how comparable* the different samples are on many important dimensions, beyond conjectured incentives (p. 418, emphasis in original).

MGV addresses all of Rogers' (1997b) concerns about differential prevalence designs. First of all, as demonstrated by the computer simulation, knowledge of *who* is malingering is actually an irrelevant consideration for accurate estimations of TPR and FPR. Second, if one administers a test (in addition to the test under investigation) for which TPR and FPR are well estimated, one can solve equation (2) for P+ to approximate *how many* are malingering (this is the process that is described in Figure 1, panel C). Third, the *meaning* of scores will be obvious based on a plot of the rates of malingering against the rates of scores. Fourth, one can *compare samples* on ''*important dimensions*'' by plotting the rates of demographics or other variables across comparison groups. Yet another advantage is that one can include all observations of a sample in an analysis of diagnostic efficiency, thereby increasing generalizability over ''known-groups'' comparisons. In CGV, or a ''known-groups'' comparison, one often eliminates ''indeterminate'' or middle

range cases (e.g., choosing to examine only the upper and lower quartiles of a range of scores). This has a potentially biasing effect and may limit interpretation of the new test to ''clear cut'' cases. Finally, because MGV does not round any and all estimations of P+ to 0 or 1 as in CGV (e.g., Rogers et al., 1998, rounded estimations of P+ from .98 to 1 and from .05 to 0 for their two mixed groups[3]), one avoids propagating known error in determining the TPR and FPR of the new test.

AN APPLICATION OF MGV IN MALINGERING DETECTION RESEARCH

The diagnostic efficiencies of the Rey 15-Item Memory Test were estimated by MGV within a large sample of individuals undergoing pre-trial mental health evaluations. Estimates of TPR and FPR for each potential score of the test were derived by MGV. With this information, receiver operating characteristic curves (ROC curves; Hanley & McNeil, 1982) were plotted in order to compare the test's diagnostic efficiencies to guessing about malingering. This process was completed three times, using different estimates of P+. The relationship between malingering and demographic variables within the sample was examined.

Participants

Participants were 723 men admitted to the U.S. Medical Center for court-referred evaluations related to criminal prosecution between November 1993 and August 1997 who completed routine psychological testing. The types of evaluation overlapped for most individuals but included 511 competency evaluations, 313 insanity evaluations, 11 risk assessments, 62 general psychological evaluations for issues related to sentencing, and 126 commitments for treatment to restore competency to stand trial. Age ranged from 18 years to 72 years ($M=36.6$, SD$=10.8$). Years of education ranged from 0 to 20 years ($M=10.7$, SD$=3.3$). Four hundred and eight were White, 196 were African–American, 81 were Hispanic, 26 were Native American, 5 were Asian–American, and 7 were from other regions around the world. Most participants ($n=667$, 92.3%) spoke English as a primary language. The most common other primary language was Spanish; some participants spoke Arabic, Swahili, Farsi, Navajo, or Dutch as a primary language. When indicated, tests were administered with the assistance of a

[3]The conclusion of Rogers and his colleagues (1998) that their ''known'' malingering group was *pure* is suspect for a reason beyond rounding error. They compounded an error of Rogers, Bagby, and Dickens (1992), who intimated that the PPP of the Structured Interview of Reported Symptoms (SIRS) in its validation sample was a stable characteristic of the test (Table 16, p. 24). The PPP for three or more ''probable malingering'' scores (the recommended cutoff to predict malingering) was based on a prevalence of malingering of just over 50% (206/403) in the validation sample. Consequently, the PPP will be lower for other samples with lower rates of malingering and higher rates of honest responding. Based on the reported TPR$=0.485$ and FPR$=0.005$ (p. 24), the PPP for three or more positive scores in their validation sample should have been reported as 0.99 (100 true positives divided by 101 positive scores). However, at a base rate of 5% malingering of psychotic symptoms in a clinical sample, the PPP of three or more positive scores drops to about 0.83. Consequently, it is possible that the ''known'' malingering group's membership in Rogers et al. (1998) actually includes upwards of 10% to 20% false positives.

translator. Most participants were literate ($n=638$, 88.2%); 67 (9.3%) reported they were illiterate; 18 (2.5%) claimed to be barely literate.

Instruments

Rey 15-Item Memory Test (RMT)

A commonly administered malingering test, the RMT (Rey, 1958), is a visual recall memory task that comprises five rows of three related items (e.g., 1, 2, 3; circle, square, triangle). Defendants were asked to study the stimulus items for 10 seconds. After a delay of 10 seconds, they were asked to write down as many of the items as they could remember, in the same order as presented, on a blank sheet of paper. The structure of the RMT is intended to aid recall of the stimulus items (Bernard, 1990). A generally accepted cut score to predict malingering is the reproduction of only eight or fewer items (Bernard & Fowler, 1986; Lezak, 1983; Rey, 1958). Lee, Loring, and Martin (1992) suggested that this cut score was too nonspecific, incorrectly identifying 7 of 100 temporal lobe epilepsy patients. Morgan (1991) and Schretlen, Brandt, Krafft, and Van Gorp (1991) found instances in which persons with severe memory deficits, or other serious neurological disorders, failed to complete at least nine items on the RMT. Guilmette, Hart, Giuliano, and Leininger (1994) reported that a cutoff of seven or less incorrectly identified 8 of 20 moderately to severely brain damaged individuals and 6 of 20 depressed psychiatric inpatients (FPR=.30 to .40). Greiffenstein, Baker, and Gola (1996) concluded that a cut score of nine or less demonstrated a TPR of .64 and FPR of .26 for 55 traumatically brain injured individuals and 90 minor head injury persons claiming permanent severe disability.

Word Recognition Test (WRT)

The WRT (Rey, 1941) is another malingering test, used in this study to predict the incidence of malingering within the sample of defendants. The WRT is composed of two word lists, one of 15 words (stimulus list) and the other of 30 words (memory test). The memory test contains the 15 stimulus words and 15 distractors. For this sample, the stimulus list was read to the examinee. The memory test was then read and the examinee was instructed to say ''Yes'' if a word was recognized as being on the stimulus list and ''No'' if it was not. The score was derived by subtracting the number of misrecognized words from the number of correctly recognized words. Greiffenstein et al. (1996) reported a TPR of .72 and a FPR of .16 for a cutoff score of 4 or less.

Procedure

Estimations of the Base Rate of Malingering Within this Sample

Three methods were employed to estimate the rate of malingered cognitive impairment within this sample. The first method of estimation involved *clinical*

ratings of the probability of malingering generated by primary clinicians prior to psychological assessment. The second was an estimation of the rate of malingering based on the *proportion of positive test scores on the WRT* based on the estimates of its TPR and FPR given by Greiffenstein et al. (1996). Finally, the rate of malingering was estimated by a *Bayesian procedure* recommended by Mossman and Hart (1996) that combines information from clinical ratings and WRT test scores.

Estimations by clinical ratings. Prior to testing, clinicians generated an estimate of the probability that the defendant would feign cognitive impairment. These ratings typically were produced after a brief initial interview (about 15 to 30 minutes) and can be construed as a "hunch." The rating was in the form of a number from 0 to 100, inclusive, with low numbers representing a low likelihood of malingering. Mossman and Hart (1996) proposed such hunches as a means of estimating the pre-test likelihood of group membership of individuals. Dawes (1967) showed how valid clinical judgments are accurate estimators of clinical base rates. That is, a valid hunch that a person is 20% likely to feign cognitive impairment is equivalent to saying: "Twenty percent of individuals like this one will feign cognitive impairment" (Dawes, 1967). Ratings were averaged to generate an estimate of the base rate for an entire sample or for subsamples.

Estimates based on proportion of WRT positive test scores. Given the reported TPR and FPR for the WRT (Greiffenstein et al., 1996), and given the proportion of positive WRT test scores within a sample, the incidence of malingering was estimated by equation (2). Figure 1, panel C, presents a visual representation of this process.

Estimates derived by Bayesian procedure. Mossman and Hart (1996) reported a method for interpreting test scores in light of the "pre-test likelihood of group membership." In this case, the proportion of positive WRT test scores were interpreted in light of the clinical ratings to estimate the pre-test (i.e., pre-RMT) likelihood of the group membership of the individual. The probability of malingering given the test score was then computed by means of Bayes's theorem (Meehl & Rosen, 1955):

$$\mathrm{P(M/+)} = \frac{\mathrm{P+} \times \mathrm{TPR}}{(\mathrm{P+} \times \mathrm{TPR}) + (\mathrm{P-} \times \mathrm{FPR})}, \tag{5}$$

and

$$\mathrm{P(M/-)} = \frac{\mathrm{P+} \times \mathrm{FNR}}{(\mathrm{P+} \times \mathrm{FNR}) + (\mathrm{P-} \times \mathrm{TNR})},^{4} \tag{6}$$

[4] $P(\mathrm{M/+})$, also known as PPP, indicates the probability of malingering given a positive score (WRT ≤ 4). $P(\mathrm{M/-})$ indicates the probability of malingering given a negative score (WRT > 4). $P(\mathrm{M/-})$ represents the ratio of negative scores earned by malingerers to all negative scores. In equation (6), FNR is the "false negative rate," the rate at which individuals who are malingering earn negative scores. FNR = 1 − TPR. TNR is the "true negative rate," also referred to as specificity, the rate at which non-malingerers earn negative scores. TNR = 1 − FPR. Based on Greiffenstein et al. (1996), for the WRT, TPR = .72, FNR = .28, FPR = .16, and TNR = .84.

These values were derived for each defendant according to their performance on the WRT. P+ and P− were based on the clinical rating of the primary clinician; if the primary clinician concluded there was a 60% likelihood of malingering, then P+ = .6 and P− = .4.

In order to conduct MGV, two subgroups were formed with the condition $P_1+ \neq P_2+$. All defendants judged as less than 10% likely to malinger constituted subgroup 1 ($n=372$) and subgroup 2 comprised all individuals judged as 10% likely or more likely to malinger ($n=351$). Each of the three estimation procedures were applied to subgroup 1 and subgroup 2, resulting in three separate estimations of P_1+, P_1-, P_2+, and P_2- (see Table 5).

Determining Overall Diagnostic Efficiency of the RMT

Computing S+ for each potential RMT cut-off score. ROC curve production involves computing TPR and FPR at each potential score for a test. There are 16 potential scores for the RMT (i.e., from 0 to 15 reproduced items). ''No items reproduced'' served as the first computation point; scores of 0 reflected malingering, and scores above 0 reflected compliance. The proportion of scores at 0 were computed as S_1+. The next cutoff score was ''1 item reproduced.'' At this cutoff, scores of 1 or less were considered indicative of malingering; scores above 1 were considered indicative of compliance. The proportion of scores at 1 or below were computed as S_1+. This process continued until 16 values of S_1+ were computed for subgroup 1. Sixteen values of S_2+ were computed for subgroup 2 in the same manner (see Table 5). Given *a priori* knowledge of P_1+, P_1-, P_2+, and P_2-, these values allowed computation of TPR and FPR at each score. ROC curves were generated for

Table 5. Rate of positive test scores on the Rey 15-Item Memory Test (RMT)

RMT score	Subgroup 1 (S_1+)	Subgroup 2 (S_2+)
0	0	0.006
1	0	0.006
2	0	0.009
3	0	0.023
4	0	0.040
5	0.003	0.054
6	0.013	0.114
7	0.016	0.142
8	0.040	0.182
9	0.113	0.288
10	0.142	0.313
11	0.194	0.373
12	0.392	0.630
13	0.403	0.630
14	0.505	0.695
15	1.000	1.000

Note: RMT score is number of items reproduced. Subgroup 1, $n=372$, was considered to represent a lower rate of malingering than subgroup 2, $n=351$. Rates represent proportion of individuals in group earning that score or lower.

each estimation method by plotting TPRs against FPRs. In this study, the area under the ROC curve (AUC) reflects the probability that a compliant individual will receive a higher score on the RMT than a malingerer (Hanley & McNeil, 1982).

Results

Estimations of Sample Base Rate

Three estimations of malingered cognitive impairment were computed for the entire sample (see Table 6).

(a) The average *clinical rating* for the 723 defendants was 15.5 (SD = 20.8), estimating the average rate of malingering at .155.
(b) The rate of positive WRT test scores (WRT $\leq$ 4) for the entire sample was .163. Analysis by equation (2) indicated that this rate of positive scores was primarily accounted for by the FPR of .16 (most positive scores were generated by compliant individuals). Consequently, based on the *proportion of positive WRT scores*, equation (2) computed the rate of cognitive malingering in the entire sample at only .005.
(c) *Bayesian estimation*, evaluating WRT test scores in light of individual pre-test likelihood estimates of malingering (clinical rating), calculated the rate of malingering in the entire sample as .135 (SD = .239).

Estimations of Subgroup Base Rates

The entire sample was divided into two subgroups based on pre-test clinical ratings. Three estimations of the rate of malingered cognitive impairment in subgroups were computed (see Table 6).

(a) Individuals with ratings of less than 10% comprised subgroup 1 (n = 372, mean rating = .031, SD = .027); subgroup 2 was composed of those with ratings of 10% or higher (n = 351, mean rating = .287, SD = .235).
(b) The rate of positive WRT test scores within subgroup 1 (rate = .078) resulted in an estimation of no malingerers in subgroup 1. Within subgroup 2, the estimation of the rate of malingering was .168, based on a rate of positive WRT test of .254.
(c) Finally, the rate of cognitive malingering in each subgroup was estimated by use of Bayesian equations (5) and (6). The rate of malingering for subgroup 1 was

Table 6. Estimate rates of cognitive malingering for the total sample and for subgroups

Method of estimation	Total sample	Subgroup 1 (P_1+)	Subgroup 2 (P_2+)
Mean clinical rating	.155	.031	.287
WRT positive scores	.005	0	.168
Mean Bayesian estimation	.135	.038	.259

Note: N for total sample = 723. n for subgroup 1 = 372. n for subgroup 2 = 351. Subgroup 1 included all individuals with pre-test clinical rating of likelihood of malingering below 10%. Subgroup 2 comprised individuals whose ratings range from 10% to 100%. WRT is the Word Recognition Test.

estimated at .019 (SD = .038); the rate of malingering for subgroup 2 was estimated at .259 (SD = .295).

Subgroup 2 was consistently estimated to manifest a higher rate of malingering than subgroup 1. Note in Table 5 that subgroup 2 consistently yielded a higher rate of positive scores than subgroup 1.

Estimations of TPR and FPR for RMT

The TPR and FPR for each cutoff score for RMT was computed by means of equations (3) and (4), using the estimates for rates of malingering for subgroups 1 and 2, and observing the rate of individuals scoring at or below each RMT cutoff score (see Table 6). This process was completed three times, once for each of the three estimations of the rate of malingering. These TPRs and FPRs are reported in Table 7. At each cutoff score, the TPR was plotted against the FPR to generate an ROC curve. These curves are shown in Figure 5. Areas under the curve (AUCs) were estimated for each curve by means of the trapezoidal method of computing area and standard errors for AUCs were computed by a procedure reported by Hanley & McNeil (1982). AUCs ranged from 0.94 to 0.98 (see Table 8); no significant differences existed between curves. Based on these results, for any randomly pair of tests generated by a malingering or a cooperative test taker, the cooperative test taker will earn a higher score on the RMT 94% to 98% of the time.

Table 7. Estimated true positive rate (TPR) and false positive rate (FPR) at each potential cutoff score of the Rey 15-Item Memory Test (RMT)

Method of estimating proportion of malingerers in subgroups 1 and 2 (P_1+ and P_2+)						
	Mean clinical rating		WRT positive scores		Mean Bayesian estimation	
RMT score	TPR	FPR	TPR	FPR	TPR	FPR
0	.023	−.001	.036	.000	.025	.000
1	.023	−.001	.036	.000	.025	.000
2	.034	−.001	.054	.000	.037	−.001
3	.087	−.003	.137	.000	.094	−.002
4	.143	−.001	.223	.003	.154	.000
5	.196	−.003	.307	.003	.211	−.001
6	.395	.001	.614	.013	.426	.005
7	.493	.001	.766	.016	.531	.006
8	.577	.023	.885	.040	.620	.029
9	.775	.092	1.155	.113	.828	.099
10	.789	.121	1.160	.142	.841	.128
11	.872	.172	1.259	.192	.926	.180
12	1.293	.363	1.809	.392	1.365	.373
13	1.262	.376	1.754	.403	1.331	.385
14	1.224	.482	1.636	.505	1.282	.490
15	1.000	1.000	1.000	1.000	1.000	1.000

Note: TPR and FPR were derived by use of Dawes–Meehl (1966) equations cited in the text as (3) and (4). TPR and FPR are estimated for cutoff scores at or below the score cited. Negative probability values are truncated to 0. Probability values greater than 1 are truncated to 1.

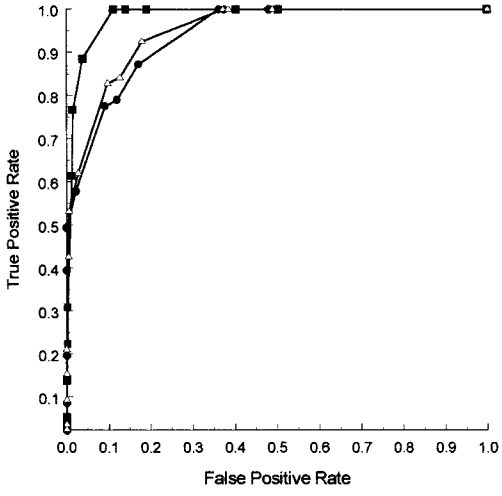

Figure 5. Receiver operating characteristic curves were plotted for the 16 values of corresponding FPR and TPR for each potential cutting score of the Rey 15-Item Memory Test. Squares represent the values of FPR and TPR predicted by Word Recognition Test estimates of the rate of malingering. Area under the curve (AUC) was equal to 0.983. Triangles represent the values generated by Bayesian estimation (AUC = 0.947); circles represent the values generated by clinical ratings (AUC = .939).

Table 8. Values of area under the curve (AUC) for receiver operating characteristic curves generated by different methods of estimating proportion of malingerers in subgroups 1 and 2 (P_1+ and P_2+)

Estimation method	AUC	Standard error
Mean clinical rating	.939	.016
WRT positive scores	.983	.047
Mean Bayesian estimation	.947	.021

Note: AUC was computed by the trapezoidal method. Standard errors were computed according to methods described by Hanley and McNeil (1982).

Demographic Variables

Within subgroup 1, the mean age was 37.6 ($n = 372$, SD = 10.7); within subgroup 2, the mean age was 35.5 ($n = 349$, SD = 10.0). The age difference between subgroups was significant ($t = 2.69$, df = 719, $p < .05$), although the effect of age appeared to be small (Cohen's $d = .21$). Mean years of education within subgroup 1 (11.3, SD = 3.2, $n = 366$) was significantly greater than the mean years of education within subgroup 2 (10.0, SD = 3.3, $n = 338$; $t = 5.33$, $p < .05$, Cohen's $d = .40$). Within subgroup 1, 63.0% were White, 25.2% were Black, 7.7% were Hispanic, and 4.1% were Native American. These proportions were significantly different from the rates in subgroup 2 (51.4% White, 30.1% Black, 15.3% Hispanic, and 3.2% Native American; chi-square = 15.2, $df = 3$, $N = 711$, $p < .05$).

MGV allows one to compare these demographics as a function of malingering within the subgroups (see Figure 6) by comparing points at which the lines cross axes at $y = 0$ and $y = 1$. For example, about 15% of the total sample of criminal defendants were 24 years old or younger (panel D of Figure 6). MGV estimates that about 10% of non-malingerers and about 33% of malingerers in this type of setting are 24 years or younger (Figure 6, Panel A). About 22% of the sample

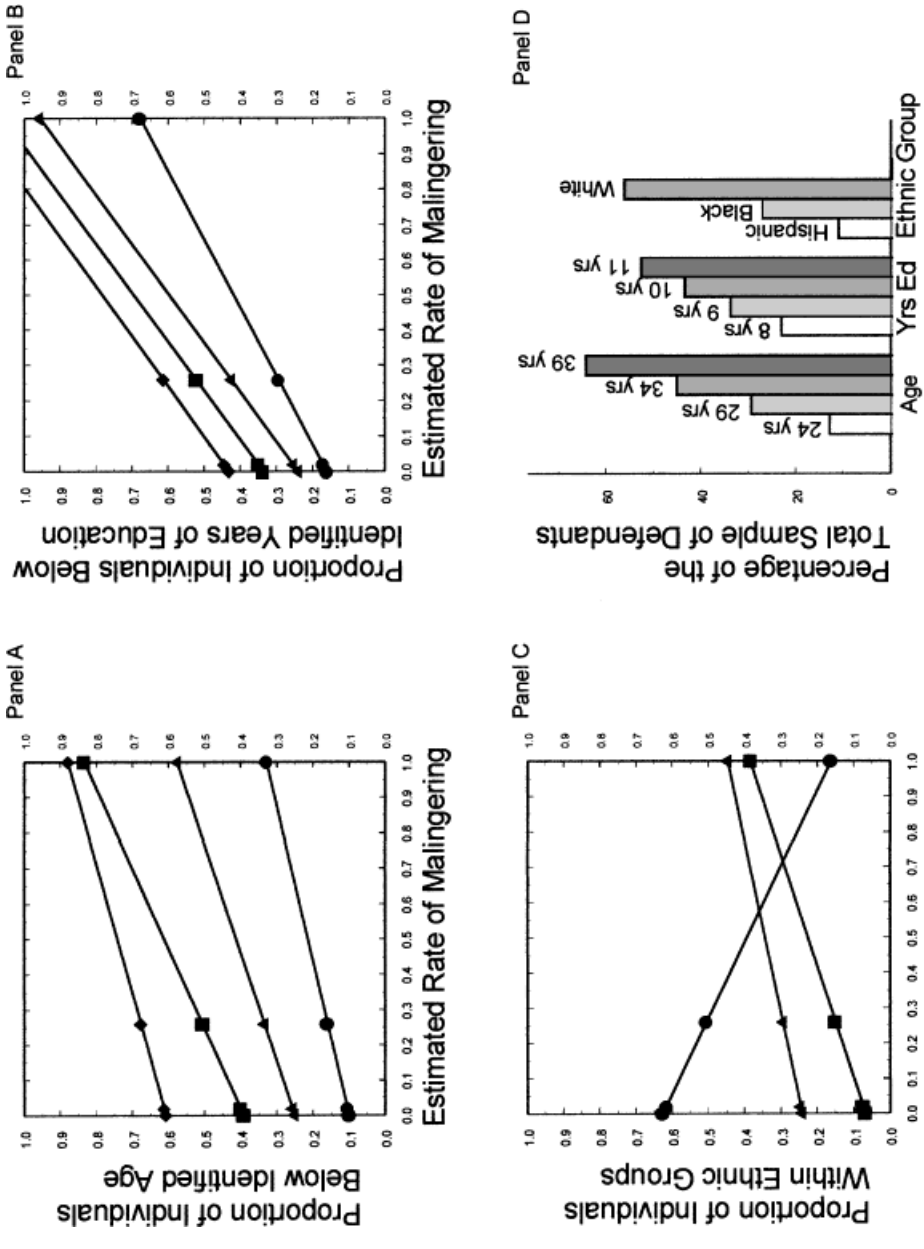

Figure 6. Rates of demographic characteristics (age, years of education, and ethnic group) were plotted against the corresponding rates of malingering generated by Bayesian estimation for subgroup 1 ($n = 372$) and subgroup 2 ($n = 351$). In panel A, the circles represent the proportion of individuals below 25 years of age, triangles represent the rate below 30 years, squares represent the rate below 35 years, and diamonds represent the rate below 40 years. In panel B, circles represent the proportion of individuals with less than nine years of education, triangles represent the rate less than 10 years, squares represent the rate less than 11 years, and diamonds represent the rate less than 12 years. In panel C, circles represent the rate of Whites in each subgroup, triangles the rate of Blacks, and squares the rate of Hispanics. Panel D shows the rate of each demographic for the total sample ($N = 723$). The bars representing age show the rate of individuals aged 24 or lower, 29 or lower, 34 or lower, and 39 or lower. The bars representing years of education show the rate of individuals with eight years of education or less, nine years or less, 10 years or less, and 11 years or less.

had eight years of education or less. MGV estimates that about 18% of non-malingerers and about 70% of malingerers have eight years of education or less (Panel B). Finally, about 55% of the total sample was White. MGV estimates that about 60% of non-malingerers and about 15% of malingerers are White (Panel C).

Discussion

The AUCs computed for the RMT were surprisingly high. A review of Table 7 suggests that a reasonably good cut-off score for the RMT is eight or fewer items reproduced. This score was associated with small FPR values (.023 to .040) and moderately-sized TPR values (.577 to .885), is consistent with Rey's (1958) recommendations, and is logically coherent. That is, scores of seven or eight result from incomplete rows; incomplete rows are unexpected, given that the sequential nature of rows aids recall (Bernard, 1990).

These values do not comport with the previously cited studies involving neuropsychological examinees, which found the RMT to demonstrate far more limited sensitivity and specificity. Table 9 shows the differences in TPR and FPR values obtained for this study and some previously cited studies. In this study, measures of TPR were consistently higher than previous studies, and measures of FPR were consistently lower. Some other potential hypotheses about the basis for these differences include: (1) real differences exist between populations of neuropsychological examinees and criminal defendant examinees in terms of the TPR and FPR of the RMT; (2) previous studies of the RMT using CGV were flawed and generated inflated FPR values and lowered TPR values; or (3) errors in estimating the rates of malingering within subgroups 1 and 2 for this study contributed to an increase in estimated TPRs and a decrease in estimate FPRs.

Differences Between Populations

There are obvious differences in the experiences of typical neuropsychological examinees and the criminal defendants represented in this sample.[5] Most neuropsychological examinees do not face the prospect of incarceration; few criminal defendants in this sample were coping with the prospect of permanent neuropsychological impairment. Neuropsychological examinees in civil litigation seem more likely to have attorneys who are more invested in knowing the nature and purpose of psychological tests administered to their clients, to the point that some coach their clients in how to take the tests (Wetter & Corrigan, 1995; Youngjohn, 1995). Most criminal defense attorneys are probably less knowledgeable about or less involved in the process of assessing cognitive capacities than attorneys involved in civil litigation on brain injury issues.

The nature of psychological test batteries for these two populations are almost certainly extensively different. Neuropsychological examinees typically are administered a wide range of procedures that assess all sorts of brain-behavior functions. Within that context, the RMT is an obviously easy test, although it is often

[5]A small number of defendants in this sample were referred to the U.S. Medical Center specifically for neuropsychological assessment.

Table 9. Comparisons of true positive rates and false positive rates for various cutting scores of the Rey 15-Item Memory Test

	Cutting score			
Study	<8 items	<9 items	<10 items	*n*
Schretlen et al. (1991)				
TPR	—	.145	.184	76
FPR	—	.270	.358	148
Lee et al. (1992)				
TPR	.375	.375	.440	16
FPR	.043	.071	.157	140
Guilmette et al. (1994)				
TPR	.050	.150	.250	20
FPR	.400	.450	.450	40
Greiffenstein et al. (1996)				
TPR	—	—	.644	90
FPR	—	—	.283	55
Current study				
TPR1	.493	.577	.775	208
FPR1	.001	.023	.092	515
TPR2	.766	.885	1.000	121
PR2	.016	.040	.113	602
TPR3	.531	.620	.828	187
FPR3	.006	.029	.099	536

Note: TPR1 and FPR1 were derived from rates of malingering generated by clinical probability judgements. TPR2 and FPR2 were based on rates of malingering estimated by the Word Recognition Test. TPR3 and FPR3 were derived from Bayesian estimation of the rates of malingering. For cited studies, *n* reflects actual numbers of criterion groups. For the current study, *n* reflects estimates of P+ and P−, based on estimation method.

presented as a difficult test of memory (e.g., Arnett et al., 1995; Rogers, Harrell, & Liff, 1993). Criminal forensic evaluations are typically not so comprehensive (Borum & Grisso, 1995). The criminal defendants in this sample typically received a battery consisting of the Shipley Institute of Living Scale (Zachary, 1986), the MMPI-2, and the Validity Indicator Profile (VIP; Frederick, 1997) in a group setting and, in an individual setting, were administered (in this order) the Rey Auditory Verbal Learning Test (AVLT; Lezak, 1995), the RMT, the Dot Counting Test (Rey, 1941), the Test of Nonverbal Intelligence (Brown, Sherbenou, & Johnsen, 1982), and the WRT. (Some Spanish-speaking individuals were not administered the AVLT.) For the defendants in this study, the RMT was never presented as an easy or difficult task. Instead, it was presented, in contradistinction to the AVLT, which immediately preceded it, as a test of *visual* memory (e.g., "Now I'm going to *show* you 15 things to remember"). In addition, defendants were told that remembering the order of item presentation was important for the RMT, although recalling the order of presentation had been inconsequential in AVLT.

These factors could account for lower TPRs for the RMT in a neuropsychological sample as opposed to a criminal defendant sample. That is, malingerers participating in neuropsychological examinations may have heightened awareness of the presence of malingering detection tests (prompted by attorneys), may identify the RMT as an easy task (given the difficulty of real memory tests), or may have been coached about the RMT's ultimate purpose. Coaching on the RMT in the presence

of financial incentive results in a lower TPR for the RMT (Frederick et al., 1994). Malingerers involved in criminal defense examinations (like the assessments performed for this sample) may have less insight about the assessment of motivation and may be more susceptible to malingering tests. They may be more willing to take risks to be seen as impaired, given the fear of incarceration (as opposed to the relatively minor prospect of gaining financially).

Actual differences in neuropsychological functioning between civil litigants and criminal defendants might also account for the lower FPR of the RMT in this criminal defendant sample. As cited earlier, it has often been reported that severe neuropsychological impairment often results in poorer performance on the RMT (higher FPR, see Table 9). Organic mental conditions were infrequent among criminal defendants in this sample (about 8% from 1990 through 1997); consequently, the potential contribution of such conditions to the observed FPRs in this criminal defendant sample was minimal.

Potential Flaws in CGV Research

Problems in criterion group contamination may have contributed to inflated FPR values in the cited neuropsychological studies. Most studies that reported elevated FPRs for the RMT among putatively bona fide neuropsychological patients assumed they were motivated to perform at their best level (e.g., Guilmette et al., 1994; Lee et al., 1992; Schretlen et al., 1991). Violations of this assumption always result in spuriously inflated FPRs. Furthermore, when researchers falsely assume that coached normal participants or suspected malingerers uniformly and appropriately feigned impairment, TPRs are underestimated.

Potential Errors in Estimating the Rate of Malingering

Errors in estimating the rate of pathology in mixed groups can occur for MGV as well as for CGV and may account for the high TPRs and low FPRs in this study. If the TPR and FPR values obtained for this study are in error, they involve an inflated TPR and a deflated FPR. This will occur at the greatest extent when the rate of malingering in subgroup 1 is overestimated and the rate of malingering in subgroup 2 is underestimated. The rates of estimated malingering for subgroup 1 were .000 to .038 (Table 5). These are rather small values and were not likely inflated; hence, they could not have contributed much to changes in FPR. Consequently, if FPRs were grossly underestimated in this study, then the error most likely resulted from an *underestimation* in the rate of malingering for subgroup 2. The highest value of estimation for subgroup 2, (P+=.287), was based on clinical ratings of malingering prior to testing. Raters often fail to accurately rate probabilities at the extremes, tending to overestimate low likelihoods and underestimating high likelihoods (Dawes, 1967). This produces a moderating effect on ratings and, if present, would have generated an underestimation of malingering in subgroup 2.

Table 10 shows alternate estimates of the TPR and FPR of the RMT for a cut-off score of eight or fewer items reproduced. These alternate estimates are derived from incremental increases in the estimation of malingering for subgroup 2 at .337, .387,

Table 10. Changes in predicted values of true positive rate (TPR) and false positive rate (FPR) for a cut-off score on the Rey 15-Item Memory Test if subgroup 2 estimated rate of malingering is increased

Estimated rate of malingering			
Subgroup 1	Subgroup 2	TPR	FPR
.031	.287	.577	.023
.031	.337	.490	.031
.031	.387	.427	.038
.031	.437	.379	.046

Note: These values demonstrate the potential changes in TPR and FPR if the rate of malingering for Subgroup 2 had been underestimated at .287, and the true value were higher.

and .437, and show the effect on TPR and FPR as estimations increase. Even if the rate of malingering in subgroup 2 is substantially greater than estimated, the FPR remains low. This is important to note because a very low FPR means a positive performance on the RMT is almost certainly meaningful.

Demographic Characteristics

The analysis of demographic characteristics in this study indicates that malingerers are younger and less educated than their non-malingering counterparts. These are potentially important findings if they can be verified or replicated by further research, because they support the adaptational view of malingering proposed by Rogers (1997a). The adaptational model purports malingering occurs most commonly when (1) the context of the evaluation is perceived as adversarial, (2) personal stakes are very high, and (3) no other alternative to malingering appears viable. Those with poor educational histories and with limited life experiences would seem more incapable of generating viable alternatives to malingering than older and better educated criminal defendants.

Cornell and Hawk (1989) reported a higher rate of Blacks than Whites (56.4% vs 43.6%) among identified malingerers. One potential interpretation of their finding was that ''. . .Black defendants were less trusting of the legal system and were prone to resort to malingering'' (p. 382). This interpretation represents an instance of adapting by malingering. Cornell and Hawk expressed concern that clinicians forming judgments about malingering might have unintentionally been more skeptical about the clinical presentation of Black defendants, skewing the rate at which Blacks were categorized as malingerers. In this study, the same potential bias was possible. Subgroups 1 and 2 were formed by clinician ratings. Those with ratings less than 10% were placed in subgroup 1. Those with ratings of 10% or higher were placed in subgroup 2. If clinicians were even slightly biased in rating Blacks or Hispanics (or younger or less educated individuals) higher than Whites (or older or more educated individuals) in terms of the probability of malingering, then that might have increased the rates of Blacks and Hispanics in subgroup 2 and produce the results seen in Figure 6, panel C. The relationship between ethnic group, level of education, age, and the rate of malingering is likely to be quite complex and deserves a more careful evaluation across many different geographical and clinical settings.

SUMMARY

MGV can potentially contribute much to psycholegal research, well beyond the domain of improving the detection of malingering. Given the large number of databases regarding the rates of relevant behavior and conditions (e.g., the *Sourcebook of Criminal Justice Statistics* published annually by the Bureau of Justice Statistics), MGV can facilitate research regarding the relationship between relevant behaviors and conditions without the need for generating data for individuals. MGV can improve research regarding conditions for which operational definitions prove difficult and criterion groups are subject to contamination.

As shown in this study, test signs or other predictors of group membership currently in use may have greater validity than previously suggested by inadequate CGV designs. The RMT may be a much better test than it has seemed to be in previous research. The primary limitation to direct interpretation of findings regarding the RMT in this study is the process of estimating the rates of malingering within subgroups. Methods that can more effectively estimate the rates and latent distributions of conditions for which objectively accurate classification criteria do not exist (e.g., taxometric analysis: Meehl, 1995) will prove useful in overcoming this limitation.

REFERENCES

Alf E, Abrahams NM. 1967. Mixed group validation: A critique. *Psychological Bulletin* **67**: 443–444.

Arnett PA, Hammeke TA, Schwartz L. 1995. Quantitative and qualitative performance on Rey's 15-item test in neurological patients and dissimulators. *Clinical Neuropsychologist* **9**: 17–26.

Baldessarini RJ, Finklestein S, Arana GW. 1983. The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry* **40**: 569–573.

Bernard LC. 1990. Prospects for faking believable memory deficits on neuropsychological tests and the use of incentives in simulation research. *Journal of Clinical and Experimental Neuropsychology* **12**: 715–728.

Bernard LC, Fowler W. 1986. Assessing the validity of memory complaints: Performance of brain-damaged and normal individuals on Rey's task to detect malingering. *Journal of Clinical Psychology* **46**: 432–436.

Borum R, Grisso T. 1995. Psychological test use in criminal forensic evaluations. *Professional Psychology: Research and Practice* **26**: 465–473.

Brown L, Sherbenou RJ, Johnsen SK. 1982. *Test of Nonverbal Intelligence: A Language-Free Measure of Cognitive Ability*. Pro-Ed: Austin, TX.

Butcher JN, Dahlstrom WG, Graham JR, Tellegen A, Kaemmer B. 1989. *MMPI-2: Manual for Administration and Scoring*. University of Minnesota Press: Minneapolis.

Butcher JN, Graham JR, Ben-Porath YS. 1995. Methodological problems and issues in MMPI, MMPI-2, and MMPI-A research. *Psychological Assessment* **7**: 320–329.

Cobb S, Hunt P, Harburg E. 1969. The intrafamilial transmission of rheumatoid arthritis—II. *Journal of Chronic Diseases* **22**: 203–215.

Cornell DG, Hawk GL. 1989. Clinical presentation of malingerers diagnosed by experienced forensic psychologists. *Law and Human Behavior* **13**: 375–383.

Dawes RM. 1962. A note on base rates and psychometric efficiency. *Journal of Consulting Psychology* **26**: 422–424.

Dawes RM. 1967. How clinical probability judgments may be used to validate diagnostic signs. *Journal of Clinical Psychology* **23**: 403–410.

Dawes RM, Meehl PE. 1966. Mixed-group validation: A method for determining the validity of diagnostic signs without using criterion groups. *Psychological Bulletin* **66**: 63–67.

Elwood RW. 1993. Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review* **13**: 409–419.

Frederick RI. 1997. *The Validity Indicator Profile*. NCS Assessments: Minnetonka MN.

Frederick RI, Sarfaty SD, Johnston JD, Powel J. 1994. Validation of a detector of response bias on a forced-choice test of nonverbal ability. *Neuropsychology* **8**: 118–125.
Goebel RA. 1983. Detection of faking on the Halstead–Reitan neuropsychological test battery. *Journal of Clinical Psychology* **39**: 731–742.
Goodman LA. 1953. Ecological regression and behavior in individuals. *American Sociological Review* **15**: 351–357.
Goodman LA. 1959. Some alternatives to ecological correlation. *American Journal of Sociology* **64**: 610–625.
Greiffenstein MF, Baker WJ, Gola T. 1994. Validation of malingered amnesia measures in a large clinical sample. *Psychological Assessment* **6**: 218–224.
Greiffenstein MF, Baker WJ, Gola T. 1996. Comparison of multiple scoring methods for Rey's malingered amnesia measures. *Archives of Clinical Neuropsychology* **11**: 283–293.
Greiffenstein MF, Gola T, Baker WJ. 1995. MMPI-2 validity scales versus domain specific measures in detection of factitious traumatic brain injury. *The Clinical Neuropsychologist* **9**: 230–240.
Guilmette TJ, Hart KJ, Giuliano AJ, Leininger BE. 1994. Detecting simulated memory impairment: Comparison of the Rey Fifteen-Item Test and the Hiscock Forced-Choice Procedure. *The Clinical Neuropsychologist* **8**: 283–294.
Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**: 29–36.
Knowles EE, Schroeder DA. 1990. Concurrent validity of the MacAndrew Alcoholism Scale: Mixed-group validation. *Journal of Studies on Alcohol* **51**: 257–262.
Lee GP, Loring DW, Martin RC. 1992. Rey's 15-Item Memory Test for the detection of malingering: Normative observations on patients with neurological disorders. *Psychological Assessment* **4**: 43–46.
Lezak MD. 1983. *Neuropsychological assessment* 2nd edn. New York: Oxford.
Lezak MD. 1995. *Neuropsychological assessment* 3rd edn. New York: Oxford.
Linn RL. 1967. A note on mixed group validation. *Psychological Bulletin* **67**: 378.
Meehl PE. 1995. Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist* **50**: 266–275.
Meehl PE, Rosen A. 1955. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin* **52**: 194–216.
Morgan SF. 1991. Effect of true memory impairment on a test of memory complaint validity. *Archives of Clinical Neuropsychology* **6**: 327–334.
Mossman D, Hart KJ. 1996. Presenting evidence of malingering to courts: Insights from decision theory. *Behavioral Sciences and the Law* **14**: 271–291.
Mossman D, Somoza E. 1991. Neuropsychiatric decision making: The role of disorder prevalence in diagnostic testing. *Journal of Neuropsychiatry* **3**: 84–88.
Nicholson RA, Mouton GJ, Bagby RM, Buis T, Peterson SA, Buigas RA. 1997. Utility of MMPI-2 indicators of response distortion: Receiver operating characteristic analysis. *Psychological Assessment* **9**: 471–479.
Rey A. 1941. L'examen psychologie dans les cas d'encephalopathie traumatique. *Archives de Psychologie* **28**: 286–340.
Rey A. 1958. *L'Examen Clinique en Psychologie*. Presses Universitaires de France: Paris.
Rogers R. 1997a. Introduction. In *Clinical Assessment of Malingering and Deception* 2nd ed., Rogers R. (ed.). Guilford: New York; 398–426.
Rogers R. 1997b. Researching dissimulation. In *Clinical Assessment of Malingering and Deception* 2nd edn., Rogers R. (ed.). Guilford: New York; 398–426.
Rogers R, Bagby RM, Dickens SE. 1992. *Structured Interview of Reported Symptoms (SIRS): Professional Manual*. Psychological Assessment Resources: Odessa, FL.
Rogers R, Harrell EH, Liff CD. 1993. Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review* **13**: 255–274.
Rogers R, Sewell KW, Cruise KR, Wang EW, Ustad KL. 1998. The PAI and feigning: A cautionary note on its use in forensic–correctional settings. *Assessment* **5**: 399–405.
Rorer LG, Dawes RM. 1982. A base-rate bootstrap. *Journal of Consulting and Clinical Psychology* **50**: 419–425.
Schretlen D, Brandt J, Krafft L, Van Gorp W. 1991. Some caveats in using the Rey 15-Item Memory Test to detect malingered amnesia. *Psychological Assessment* **3**: 667–672.
Viglione DJ, Fals-Stewart W, Moxham Z. 1995. Maximizing internal and external validity in MMPI malingering research: A study of a military population. *Journal of Personality Assessment* **55**: 280–295.
Wetter MW, Corrigan SK. 1995. Providing information to clients about psychological tests: A survey of attorney's and law students' attitudes. *Professional Psychology: Research and Practice* **26**: 474–477.
Youngjohn JR. 1995. Confirmed attorney coaching prior to neuropsychological evaluation. *Assessment* **2**: 279–283.
Zachary RA. 1986. *Shipley Institute of Living Scale: Revised Manual*. Western Psychological Services: Los Angeles.

