Neuropsychology
1994, Vol. 8, No. 1, 118–125

0894-4105/94/$3.00

Validation of a Detector of Response Bias on a Forced-Choice Test of Nonverbal Ability

Richard I. Frederick, Stephen D. Sarfaty, J. Dennis Johnston, and Jeffrey Powel

Validation studies of a 2-alternative forced-choice test used to detect faked cognitive impairment are reported. In Study 1, 177 college students were given substantial financial incentives to fake believable impairment. Rate of detection was compromised by financial incentives, but the test demonstrated superior specificity and sensitivity relative to other measures of response bias. Study 2 and Study 3 included neuropsychology (n = 134) and forensic (n = 18) evaluees, who were administered several tests of response bias. One decision rule for the forced-choice test demonstrated greater sensitivity than the other measures of response bias and displayed moderately good agreement with clinician ratings of individual testing response style. Although we originally intended to identify malingering, we conclude that psychometric tasks can only detect biased responding and that the determination of malingering must be based on overall clinical evaluation.

Frederick and Foster (1991) described a two-alternative forced-choice test of nonverbal ability (FCTNV) that generated several detectors of dissimulated intellectual impairment or "fake bad" responding. In two analog studies (in which only extra credit was used as an inducement to follow experimental instructions), these measures demonstrated excellent sensitivity (correct detection) and excellent specificity (correct rejection) in categorizing college students who were asked to suppress their true cognitive ability or to give their best effort on the FCTNV. In a third simulation study with equivalent inducements, some subjects who were asked to fake bad were also informed of effective strategies for evading detection, but the sensitivity of the most effective detector was only slightly compromised.

Providing subjects with only low-level inducements to fake bad in analog studies may have generated an overestimation of the effectiveness of these measures. To accurately evaluate the robustness of a measure of response bias, one must conduct an analog study that provides realistic incentives, or one must sample the actual populations of concern (Gillis, Rogers, & Bagby, 1991; Rogers, 1988b). This article reports both an analog study and two studies within the populations of concern. In Study 1, college students were administered the FCTNV and other tests sometimes used to detect response bias in intellectual and memory testing. Financial incentives were offered to those participants who could convincingly suppress their actual cognitive abilities. In the second and third studies, we administered these tests to samples of neuropsychology and forensic evaluees, which were assumed to include some unknown proportion of persons who misrepresented their true cognitive abilities to secure some secondary gain.

Richard I. Frederick, Department of Psychology, Whiting Forensic Institute, Middletown, Connecticut; Stephen D. Sarfaty and J. Dennis Johnston, Comprehensive Neuropsychological Services, Cheshire, Connecticut; Jeffrey Powel, Behavioral Medicine Northwest, Seattle, Washington.

Study 1 of this article was funded in part by a grant from the American Psychology–Law Society.

We thank Bradley Waite and Fred Maxwell for their help in obtaining volunteers for Study 1. We also thank Marcus Sharpe, Harold Maphet, Beth Curry, and Michael Greenleaf for contributions to Studies 2 and 3, and we thank Robert Denney and three anonymous reviewers for reviewing a draft of this article.

Correspondence concerning this article should be addressed to Richard I. Frederick, who is now at the Department of Psychology, U.S. Medical Center for Federal Prisoners, Springfield, Missouri 65808.

Forced-Choice Test of Nonverbal Ability

The FCTNV, which was fully described by Frederick and Foster (1991), is a modification of the Test of Nonverbal Intelligence (TONI; Brown, Sherbenou, & Johnsen, 1982). The TONI is a picture matrix test that comprises two equivalent forms of 50 items each. For each picture puzzle, four or six answer choices are available. In the TONI, either form is presented, in order of difficulty, until the evaluee's ceiling is reached. In the FCTNV, all 100 items are presented but in a random order of difficulty. Only the correct choice (target) and one distractor are presented for each of the 100 TONI picture puzzles. This results in 100 trials of a two-alternative forced-choice task.

As in symptom validity testing (Pankratz, 1979), the number of items answered correctly serves as a measure of response bias. With alpha set to .05 (two-tailed), the range of random responding for 100 items is 42–58. Because random responding characterizes the performance of individuals with no ability at all, scores below 42 indicate that responding is biased toward incorrect answers. Although a score below 42 is inherently specific to biased responding, it was not sensitive in Frederick and Foster's (1991) study. Only a few malingerers scored below the range of random responding.

Although symptom validity testing relies only on the score as a measure of biased responding, Frederick and Foster (1991) identified several other measures that can be generated with the two-alternative forced-choice response format:

1. *Slope.* A curve describing test performance from the

least to most difficult test item is plotted for each subject. (In Frederick & Foster's [1991] study, item difficulty was determined post hoc.) An improvement in performance as the test items increase in difficulty (an unlikely event for compliant responders) results in a positively sloped performance curve. Positively sloped performance curves were completely specific to simulating malingerers but demonstrated only moderate sensitivity.

2. *Consistency ratio.* Fifty pairs of picture puzzles of equivalent difficulty (equivalent item pairs) are created, and a measure of consistency is thereby derived. The consistency ratio (CR) is defined as the ratio of (a) the number of equivalent item pairs in which both items are answered correctly to (b) the maximum possible number of equivalent item pairs in which both items can be answered correctly (which is the total test score divided by 2). If a is the test score and b represents the number of equivalent item in which both items are answered correctly, the equation for computing the CR is as follows: $CR = b/(a/2)$, or $CR = 2b/a$. In Frederick and Foster's (1991) study, CRs of .50 or less (the cutoff established a priori) were completely specific to, but only moderately sensitive to, simulating malingerers.

3. *Slope*consistency ratio.* Frederick and Foster (1991) found a significant interaction effect between slope and CR. Consequently, they generated a decision rule incorporating the product of the previous two measures (slope*CR) on the basis of post hoc analysis: "If the score is below average, and slope*CR is greater than −.0041, then suspect biased responding." The new decision rule was highly specific (a 97.1% correct rejection rate for normal controls) and highly sensitive (94.2% correct detection) to simulating malingerers who scored below average. This decision rule maintained its rate of detection in the two subsequent analog studies described by Frederick and Foster. In those three studies, below average performance was defined as any score below the mean for normal controls, that is, as any score lower than 83. For the current experiments, we restricted the range of below-average performers to those who scored at least one standard deviation below the mean. Now that 267 normal control college students have participated (in the previous study and in this one) and a more stable range of scores has been generated ($M = 84.9$, $SD = 8.1$), this decision rule for biased responding has been modified to the following: "If the score is less than 77 and the slope*CR product is greater than −.0041, then suspect biased responding." A reanalysis of Frederick and Foster's data did not result in any significant change in sensitivity or specificity.

The slope*CR decision rule appears to work because it simultaneously captures information about both performance reliability (slope) and performance consistency (CR). Even low performers should demonstrate reliable responding; that is, when they miss an item at a lower level of difficulty, they should not perform better at a higher level of difficulty. And, even poor performers should consistently answer items correctly at their highest level of ability.

4. *Pearson product–moment correlation.* The correlation between test performance and item difficulty (CORR) was not examined by Frederick and Foster (1991). Because CORR contributes to the computation of the performance curve slope, its use as a detector was presumed to be redundant. Post hoc analysis failed to support that assumption. Compared with the decision rule for slope alone, the use of CORR increased the range of sensitivity for cases of random or irrelevant responding. That is, some performance curves that result from random responding have a negative, but not significantly negative, slope. Consequently, those instances of random responding are not identified as biased by the slope decision rule (slope > 0). To use CORR as a detector, the confidence interval for random responding when alpha = .01 is computed. The negative boundary of this confidence interval is −.275, which results in this decision rule: "If CORR is greater than −.275, then suspect biased responding." A reanalysis of Frederick and Foster's data revealed that using this cutoff correctly identified 83 of 172 (48.3%) naive simulating malingerers (an increase of 17) and 6 of 56 (10.7%) informed simulating malingerers (an increase of 5).

Other Measures of Response Bias Used in the Present Studies

Other tests have been reported to detect response bias in evaluees. The following were included in one or more of the studies described in this article:

1. *Rey Memory Test* (RMT; Bernard & Fowler, 1990; Goldberg & Miller, 1986; Lee, Loring, & Martin, 1992; Lezak, 1983; and Schretlen, Brandt, Krafft, & Van Gorp, 1991). Five rows of three related items are presented visually: two variations of "1 2 3" and "A B C" each, and a circle, square, and triangle. Although only two truly distinct concepts exist (sequence and category), subjects are told to remember and reproduce "all 15 [emphasized] items." Even most moderately impaired individuals reproduce at least nine items and three rows after a brief exposure (Goldberg & Miller, but compare with Schretlen et al. and Lee et al.). Performance below that level is generally considered biased.

2. *Portland Digit Recognition Test* (Binder, 1990, 1993; Binder & Willis, 1991). This test is based on symptom validity testing and is modeled after a task described by Hiscock and Hiscock (1989). Subjects are verbally presented with a five-digit number and then asked to count backward aloud until interrupted after 5 s. They are then presented with a card showing the target number and a distractor and asked to identify the number to be remembered. After the first 18 trials, the response delay increases to 15 s; after the next 18 trials, the response delay increases to 30 s for the last 36 trials. A score of less than 27 correct (at $p < .05$, one-tailed) indicates persons who have suppressed their memory ability. In the studies reported in this article, we used a score of 26 or less as the indicator of biased responding; Binder (1993) suggested a higher, empirically based cutoff.

3. *Dot Counting Test* (DCT; Lezak, 1983; Paul, Franzen, Cohen, & Fremouw, in press). Twelve cards with dots arranged in either organized or random patterns ($n = 6$ for each type of card) are presented individually to subjects, who are instructed to count the dots as quickly as possible without making mistakes. The total counting time for organized patterns is expected to be far less than that for the random patterns.

4. *Word recognition versus word recall* (WORD; Lezak, 1983). A list of 15 words are presented to subjects from the Word Recognition Test (WRT; Lezak, 1983), which comprises 15 stimulus words followed by 30 recognition word choices and the first trial of 15 words from the Auditory Verbal Learning Test (AVLT; Lezak, 1983). In Study 2, Johnston and Powel used the WRT (revised to contain 16 stimulus and 32 recognition words) and the first trial (16 words) of the California Verbal Learning Test (Delis, Kramer, Kaplan, & Ober, 1987). For WORD, response bias was suspected in individuals who recalled more words than they recognized.

5. *21-Item Word List* (21WRD; Iverson, Franzen, & McCracken, 1991; Wilhelm, Franzen, & Grinvalds, 1991). Twenty-one words are presented to and recalled by evaluees (21WRD1). Immediately after the recall, 21 word pairs (each containing a target word from the original list) are presented verbally. The evaluee must identify (recognize) one word in each pair as a word from the original list (21WRD2). According to the decision rule established by Iverson et al., recall of fewer than 3 words or recognition of fewer than 13 words indicates a biased performance.

6. *Memorization of 16 Items* (MSIT; Paul et al., in press; Wilhelm et al., 1991). This is a modification of the RMT. No geometric figures appear. There are four rows of four items, for example, "1 2 3 4." The criterion for suspected response bias is reproduction of fewer than two complete rows.

7. *Subjective impression of response style.* Rogers (1988a, pp. 4–5) identified six types of response style: malingering, defensive, irrelevant, random, honest, or hybrid. Hybrid responding refers to the combination of any of the other five categories. On the basis of a comprehensive evaluation, neuropsychology evaluees (in Study 2) were assigned to one of these categories by their examining neuropsychologist.

Study 1: Offering Financial Incentives and Administering Other Tests of Malingering to College Students

In Frederick and Foster's (1991) study, no substantial external motivation to fake bad existed, as all subjects could realistically expect to earn extra credit simply by showing up for the experiment. In the present study, financial incentives were offered to those subjects who could both suppress their cognitive abilities and avoid detection. Study 1 also compared the effectiveness of other measures of response bias to the slope*CR decision rule.

Method

Subjects. Subjects were 269 introductory psychology students (118 men and 151 women). Their mean age was 22.1 years (SD = 6.9). Two hundred twenty-five were White; 21 were African American; 6 were Hispanic; 4 were Native American, 7 were Asian, and 6 identified themselves as "other."

Procedure. The amount of money to be offered was determined in a pilot study of college students (N = 82; 34 men and 48 women). The modal response (n = 18) for "substantial incentive" was $20; 58% indicated that a bonus of $20 or less would provide significant motivation to appear realistically impaired. After a consideration of available funds, the anticipated low likelihood of having to make a payment, and the possibility that subjects in the pilot study had inflated (but had certainly not lowballed) the amount that would motivate them, we chose $20 as the bonus amount.

Group testings (maximum number per group = 12, time of participation = approximately 105 min) were conducted independently in college classrooms. The following tests were administered (in order): AVLT, RMT, (another word and memory test not described here, time = 35 min), WRT, MSIT, 21WRD, FCTNV, and DCT. All tests but the FCTNV and the DCT were presented by means of a videotape and monitor to ensure consistent articulation and time of exposure. The FCTNV was distributed after the video presentations. While students completed the FCTNV, each was briefly escorted to another testing area where the DCT was administered. For the DCT, students were asked to say "stop" when the count was derived and then write their answer down. The WRT was modified from Lezak (1983) to preclude the repetition of identical stimulus words on the AVLT and the 21WRD. Because tests were administered in groups of 12 or fewer, all word recall tests were modified by having the subjects write down their responses rather than responding verbally. On word recognition tests, the recognition items were printed, and subjects identified their choices as the words were read from the videotape. On the 21WRD, the recall data sheets were collected prior to the recognition test (21WRD2).

All subjects provided informed consent prior to participation, and all received extra course credit for their participation. *Compliers* (n = 92) were asked to take all tests and to give their best effort. They were informed that 2 of all the subjects who were identified as compliant would be chosen at random to receive a bonus of $20. The other subjects were asked to fake bad and were told that each subject who avoided detection on all the tests would receive a bonus of $20. Because their biased responding was to be deliberate and directed toward financial gain, they were referred to as *malingerers.*

Malingering subjects read a scenario (slightly amended from Iverson et al., 1991, p. 671) that described a believable situation in which a person would be motivated to fake brain and memory impairment. As Rogers (1988b) suggested, they were given about 10 min to read the scenario and contemplate possible ways to fake believable impairment. Naive malingerers (n = 94) were asked to fake believable brain impairment without being obvious. Informed malingerers (n = 83) were also given a "cheat sheet" that described effective strategies to fake believable impairment and to avoid detection for each type of test to be administered. The cheat sheet also cautioned them to appear attentive, involved, and motivated. Prior to each test, the test instructions were read, informed subjects were directed to reread identified pertinent paragraphs of the cheat sheet, and then the instructions were repeated. Following the administration of tests, subjects completed a questionnaire on which they rated their motivation and incentive to comply with experimental instructions and their judgment about the likelihood that they would receive a bonus for their participation.

Results

All of the compliers reported that they gave their best effort on most or all of the tests they were administered. One hundred thirty (73.4%) of the malingering subjects reported that they gave their best effort to fake believable brain impairment on most or all of the tests. Eleven malingering subjects (6.2%) reported that they gave little or no effort to fake believable brain impairment, and 36 (20.3%) reported that they gave some effort but not their best effort. When asked how

much the $20 bonus served as an incentive to fake believable brain damage, 32 (18.1%) responded "none," 94 (53.1%) responded "a little," and 51 (28.8%) responded "a lot."

The percentages of malingering subjects detected with the slope*CR decision rule were computed. Malingerers who obtained average or better scores were not included in this computation, as they were considered not to have met the criterion for producing believable impairment. This resulted in the deletion of 12 naive subjects and 26 informed subjects from the computation. The remaining subjects were referred to as low functioning. With the incentive of $20, detection was limited to 81.7% of low-functioning naive subjects and 49.1% of low-functioning informed subjects. Using the procedures described in Fleiss (1981), we compared these rates with detection rates for low-functioning subjects from Frederick and Foster's (1991) study (no-bonus naive = 94.2%; no-bonus informed = 74.1%). No significant interaction occurred between information about detection methods and the presence of financial incentives (z = 1.245). Main effects for incentives were significant (naive subjects, z = 2.91; informed subjects, z = 2.51). Main effects for coaching were significant (no incentives, z = 3.94; incentives, z = 3.88; see Table 1 and Table 2).

Figure 1 shows the effectiveness of each test as a detector of response bias within each condition. Sensitivity for any test is indicated by the percentage of malingerers identified, and specificity is derived by subtracting the percentage of compliers identified by a test as biased responders from 100%. (To make consistent comparisons across types of test, we included subjects who scored in the average or better range on the FCTNV in the data for the slope*CR decision rule in Figure 1.) The DCT and WORD were poor detectors and had false-positive rates roughly equivalent to hit rates. All individuals identified by the MSIT and 21WRD1 were also detected by the RMT and 21WRD2, respectively.

The RMT, 21WRD2, CR, and CORR all demonstrated excellent specificity but only limited sensitivity. Nevertheless, they were the best measures with which to compare the detection rate of the slope*CR decision rule. Thus, the overall detection rates for RMT, 21WRD2, CR, and CORR were collapsed into one category called *other measures*. Table 3 shows the proportion of low-functioning subjects (n = 154) identified as biased responders by the slope*CR decision rule (proportion = .64) and by other measures (overall proportion = .55). The level of agreement between these two detection methods for low-functioning subjects in all categories was moderately good (κ = .45, z = 5.70). When only naive malingerers (n = 94) were considered, the proportion detected by the slope*CR decision rule was .71 and by other measures was .69 (κ = .49, z = 4.77).

Table 1
Proportion of Malingerers Detected in Each Group

Group	n	Proportion detected
Naive		
No bonus	172	.942
Bonus	82	.817
Informed		
No bonus	54	.741
Bonus	57	.491

Table 2
Main Effects of Financial Incentives and Information About Strategies for Avoiding Detection

Type of contrast	P1	P2	n'	z	α	$1-\beta$
No bonus (P1) vs. bonus (P2)						
Naive	.942	.817	111	2.91[a]	.0125	.682
Informed	.741	.491	56	2.51[a]	.0125	.774
Naive (P1) vs. informed (P2)						
No bonus	.942	.741	82	3.94[a]	.0125	.956
Bonus	.817	.491	67	3.88[a]	.0125	.913

Note. P1 and P2 are proportions of detected malingerers respective to groups. n' is the harmonic mean of $n1$ and $n2$. Overall alpha is .05.
[a] Greater than cutoff value at α = .0125.

Discussion

Several researchers (Bernard, 1990, 1991; Wilhelm et al., 1991) have found no significant differences between groups of malingerers who were or were not given financial incentives. A flaw in the design of all three of these studies was the low number of subjects in each comparison group (30 or fewer), which resulted in low experimental power. That is, the designs did not allow a sufficient opportunity to find any

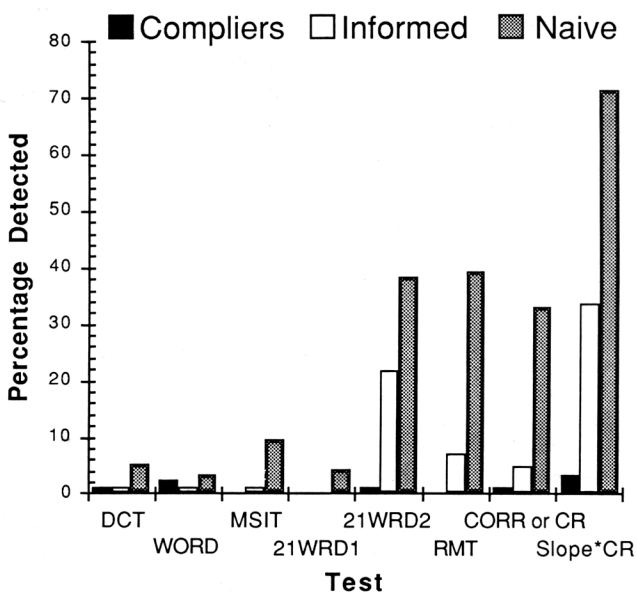

Figure 1. Percentage of college students detected as biased responders (out of 92 compliers, 83 informed subjects, and 94 naive subjects; DCT = Dot Counting Test; WORD = comparison of word recognition and word recall; MSIT = Memorization of 16 Items; 21WRD1 = recall of 21-Item Word List; 21WRD2 = recognition of 21-Item Word List; RMT = Rey Memory Test; CORR = correlation between test performance and item difficulty; CR = consistency ratio).

Table 3
*Agreement Between Proportions of Low-Functioning Subjects Detected as Biased Responders by the Slope*CR Decision Rule and by Other Measures*

	Other measures		
Slope*CR decision rule	Detected	Not detected	Total
Detected	.461	.175	.636
Not detected	.091	.273	.364
Total	.552	.448	1.000

Note. Proportion observed = .734, proportion expected = .514, κ = .452, *SE* κ = .079, z = 5.696. CR = consistency ratio.

real difference that may have existed. Cohen (1977) recommended a power of no less than .80 to satisfy the requirement that a design is sufficient to detect a false null hypothesis. The power for each comparison in the present study was near to or exceeded that standard, which may explain why a difference was observed (see Table 2).

In this context, the slope*CR decision rule was a valid detector of suppressed performance. But a problem related to the use of incentives in an analog study is whether such findings can be replicated in a real-world setting. Gillis et al. (1991) found that the *M* test (Beaber, Marston, Michelli, & Mills, 1985) detected 79.5% of simulators but only 40% of suspected malingerers. They concluded that these differences were, in part, due to a lack of "more realistic conditions" and "higher intrinsic motivation of [actual] subjects to avoid being detected" (Gillis et al., 1991, pp. 138–139). Given that a further analysis (Rogers, Bagby, & Gillis, 1992) accurately detected 81% of the same group of suspected malingerers (n = 21), Gillis et al.'s conclusion did not remain supported. Nonetheless, our Study 1 demonstrates that financial incentives decrease sensitivity in the detection of response bias. Therefore, we conducted studies in neuropsychology clinics and forensic evaluation sites.

Study 2: Neuropsychology Evaluees

Method

Subjects. Subjects were 134 individuals (81 men and 53 women). Their mean age was 40.6 years (*SD* = 14.0, range = 15 to 75), and the mean level of education was 12.8 years (*SD* = 3.0, range = 3–20). Racial information was not collected. Subjects presented at Comprehensive Neuropsychological Services in Cheshire, Connecticut, or at Behavioral Medicine Northwest in Seattle, Washington, for diverse types of evaluations. Subjects were divided into clinical (n = 70) and nonclinical (n = 64) groups on the basis of referral source and questions. The term *nonclinical* denotes that the evaluee was referred for other than purely clinical reasons and implies that some motivation to exaggerate symptoms may have existed. Nonclinical evaluees comprised those referred for insurance medical examinations (n = 25), examinations required for criminal (n = 6) or civil (n = 27) litigation, and examinations required for disability insurance or workers' compensation (n = 6).

Procedure. As part of the routine initial interview in both clinics, evaluees were informed that their motivation to perform to their maximum ability would be evaluated and that some of the tests that would be administered might actually measure something other than what they appeared to measure. All evaluees in this study agreed to allow their test results to be used as research data, provided that individual confidentiality was maintained. All evaluees at the two clinics were routinely administered the FCTNV. Other tests of response bias (RMT, PDRT, DCT, and WORD) were used, but not all of these other tests were administered to all subjects. Ratings of response style were given by the neuropsychologists after consideration of the testing, interview, and historical data. That is, ratings of response style were based on the comprehensive evaluation, with attention to reason for referral, personality issues, cultural factors, and test performance, and were not based exclusively on cutoff scores. Because of changes in design, only 95 evaluees received a rating.

Results

Figure 2 shows the percentage of clients who were identified as biased responders by the tests used in this study. Again, not all evaluees are represented for each category. Fifteen subjects (11.2% overall; 6 clinical subjects and 9 nonclinical subjects) received positive scores on the RMT, CR, or CORR (measures that were shown in Study 1 to be completely specific to biased responders) or on the Portland Digit Recognition Test, which as we used it is inherently specific. Twenty-four subjects (15.7% overall; 8 clinical and 16 nonclinical) were categorized as biased responders by the

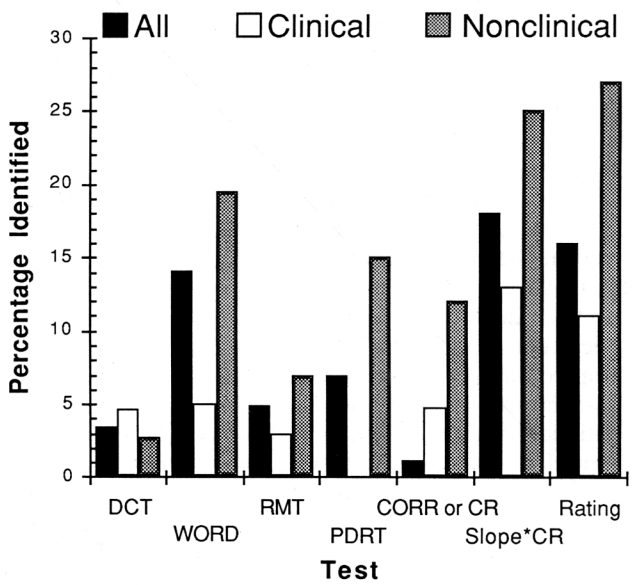

Figure 2. Percentage of 134 neuropsychology evaluees (70 clinical and 64 nonclinical) identified as biased responders. (Except for CORR, CR, and Slope*CR, numbers of evaluees for each type of test are not equal to overall group size. For the Dot Counting Test [DCT], n = 48 clinical and 57 nonclinical; for the comparison of word recognition and word recall [WORD], n = 20 clinical and 36 nonclinical; for the Rey Memory Test [RMT], n = 49 clinical and 56 nonclinical; for the Portland Digit Recognition Test [PDRT], n = 26 clinical and 21 nonclinical; for the neuropsychologists' rating, n = 61 clinical and 34 nonclinical; CORR = correlation between test performance and item difficulty; CR = consistency ratio).

slope*CR decision rule. In the ratings of response style suggested by Rogers (1988a), no subject received a rating of malingering. Fifteen (15.7%) subjects were assigned to the category of hybrid responding (6 clinical subjects and 9 nonclinical subjects).

The identification rates for the RMT, PDRT, CR, and CORR were collapsed into the other measures category and compared with the identification rate for the slope*CR decision rule (n = 134, κ = .55, z = 6.66). The identification rate for the slope*CR decision rule was also compared with the final rating by neuropsychologists (n = 95, κ = .42, z = 4.12).

Discussion

According to the slope*CR decision rule, about 25% of the nonclinical evaluees and about 10% of the clinical evaluees responded in a biased fashion on the FCTNV. Unfortunately, the true proportion of biased responders could not be determined. But, an estimate of the minimum proportion of biased responders is provided by the detection rates for other measures (14% of the nonclinical evaluees and 8.5% of the clinical evaluees). This is an impressive finding and suggests that neuropsychologists need to attend closely to the validity of psychological test performance for their clients. Similar findings were obtained by Binder (1993), who used higher cutoff values for the Portland Digit Recognition Test than we used in this study and found exaggerated deficits in two groups of evaluees with potential financial compensation (18% of patients with brain dysfunction and 33% of mild head trauma patients). Binder, Villanueva, Howieson, and Moore (1993), using a recognition score of less than 6 on the AVLT (Lezak, 1983), concluded that approximately 5% of clinical neuropsychology evaluees and about 26% of nonclinical neuropsychology evaluees were properly classified as poorly motivated to perform. These findings suggest that the proportions generated by the slope*CR decision rule are reasonable. Further evidence is provided by the comparison of the proportion of ratings of hybrid responding by the neuropsychologists to the almost identical proportion of evaluees identified as biased responders by the slope*CR decision rule (see Figure 2). It seems fair to conclude that Study 2 successfully cross-validated the slope*CR decision rule in a more realistic condition.

Study 3: Forensic Evaluees

Method

The third study was conducted over a 3-year period primarily at the Whiting Forensic Institute, a maximum-security psychiatric hospital in Middletown, Connecticut. Seventeen male forensic evaluees (mean age = 31.0 years, SD = 9.2; 5 White, 7 African American, and 5 Hispanic) were tested. Fifteen were a subgroup of all patients referred by the courts for an evaluation of their competency to stand trial. The other 2 subjects were a presentence evaluee and a prerelease evaluee. Four of the competency evaluees were referred from regional state hospitals. Evaluees were limited to individuals who were referred by their clinical teams specifically for an evaluation of faked cognitive impairment. These persons had been noted to be uncooperative with the ongoing evaluation and to have unusual or contradictory cognitive symptoms. Before consenting to the testing, evaluees were informed that test results would be reported to the clinical team and would be available to the court that ordered the evaluation. Reports to clinical teams described test-taking behavior, provided test scores, and included an admonishment to use findings only to make decisions about the validity of psychological testing for that individual.

The assessment battery for evaluees comprised the FCTNV, RMT, DCT, WRT, and AVLT, but 3 evaluees did not receive the DCT, one did not receive the WRT, and one did not receive the RMT. One forensic evaluee was tested twice, after being discharged and readmitted; his two evaluations occurred 4 months apart. The total number of evaluations was 18.

A group of controls was also tested (n = 18; mean age = 35.6 years, SD = 6.8; 14 were White and 4 were African American). These individuals had been found not guilty by reason of insanity and ordered to undergo inpatient psychiatric treatment at the same hospital as most of the competency evaluees. They were volunteers for this study and participated with the agreement that test results would be anonymous and would not be entered into their medical records. By policy, they could not be individually compensated for their participation, but the wards on which they resided received doughnuts and coffee. They received the same battery as the forensic evaluees, with the addition of the 21WRD.

Results

Figure 3 shows the percentage of subjects each test identified as a biased responder. In this study, other measures refers to the proportion of subjects who had a positive finding on the RMT, CR, CORR, or 21WRD2. For 14 evaluees (77.8%), the slope*CR decision rule was positive.

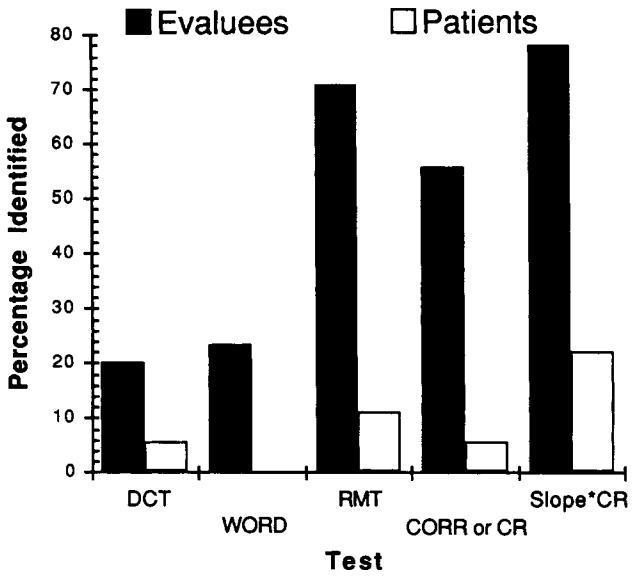

Figure 3. Percentage of forensic criminal evaluees (n = 18) and patients sentenced to psychiatric treatment (n = 18) identified as biased responders (DCT = Dot Counting Test; WORD = comparison of word recognition and word recall; RMT = Rey Memory Test; CORR = correlation between test performance and item difficulty; CR = consistency ratio).

None of those 14 individuals received a diagnosis of an Axis I disorder other than substance abuse, and all received a V-code diagnosis of malingering. Four of the evaluees (including the presentencing and prerelease evaluees) were not identified as biased responders by the slope*CR decision rule or other measures. Only one of those 4 individuals was identified as a biased responder by any of the other tests—a positive finding on the DCT. None of the 4 was labeled as malingering in discharge diagnoses. Three were diagnosed as having personality disorders; one received a diagnosis of dysthymia disorder. Obviously, this yields a κ of 1.00 for any comparison among those three indicators (slope*CR decision rule, other measures, and diagnosis) for evaluees. A comparison of the slope*CR decision rule with other measures for all 36 subjects yielded an excellent correlation ($\kappa = .889$, $z = 5.33$).

Four of the control patients were identified as biased responders by the slope*CR decision rule. But 3 of those 4 patients also had positive findings on either the CR, RMT, or 21WRD2, which suggests that they may not have been incorrectly identified by the slope*CR decision rule.

Discussion

The fact that evaluees were specifically referred for a malingering evaluation and that treatment teams had access to testing data in making their diagnoses confounded the correlation of team ratings with the slope*CR decision rule. However, each treatment team was informed that results of the evaluation had direct applicability only to the utility of conducting further psychological testing. Each report contained a caution not to use the test results in isolation but to follow standard nosology in deriving conclusions about the presence of malingering. The correlation of other measures with the slope*CR decision rule was not confounded by this design and suggested that the slope*CR decision rule was not generating false-positive or false-negative findings. As in Study 2, it seems fair to conclude that the slope*CR decision rule was validated as a measure of response bias in a realistic condition.

General Discussion

The incentive of a few dollars to a college student, the incentive of monthly income (in some instances, lifetime support), and the incentive of freedom from incarceration cannot fairly be construed as equivalent incentives for faking bad. In Aesop's fable of the fox and the hare, the crow mocks the efforts of the fox when its prey eludes him. The fox replies, "I was running only for my supper, the hare for his life." Similarly, different motivational and incentive levels are likely represented among the various populations sampled in these studies. None of the neuropsychologists concluded that evidence of biased responding on some tests justified representing the entire effort of the individual as malingered (for a review of this issue see Pankratz & Erickson, *1990*). Although symptom embellishment was suspected in a number of cases, competing explanations for malingering could not be ruled out. Fatigue, anxiety, nonconscious motivation to have symptoms validated, and peculiarities of an individual's own concept of his or her illness appeared to influence performance.

Neuropsychology patients presented with a wide variety of complaints. These did not necessarily include endorsement of nonverbal complaints. The question remained whether patients with other types of complaints might have generated positive findings on measures of response bias that appeared more related to their complaints. Further studies, in which the tests used include those that by face validity correspond to the evaluee's complaints, are indicated. The neuropsychologists noted that some patients generally complied with psychological tests but made gross errors of omission or exaggeration in verbal self-report of age, length of coma, or findings of other providers. Further studies are needed to determine whether subjects are more willing to dissimulate on verbal self-report than on structural task performance.

The FCTNV was originally intended to serve as a malingering test. That goal now appears to have been misguided. It is unrealistic to expect a psychological test to identify malingering. Even though some psychological tests can identify biased responding, such a finding is only one component of malingering and is also a component in other conditions. An understanding of the genesis and meaning of observed response bias depends on a thorough evaluation by a qualified clinician.

Analog studies provide a cost-effective way to investigate new measures of response bias. Even if the level of motivation to successfully fake bad on a psychological test in an analog study cannot adequately compare with that in the real environment of assessment, the number of effective strategies for performing in an impaired fashion is finite, and most if not all will be revealed in an analog study. That is, even when completely motivated to fake bad, an individual is still limited by the number of available dissimulation strategies. If the cautions provided by Rogers (1988b) are heeded, one can work to design measures sensitive to response bias. A measure's effectiveness can be tested among populations of limited generalizability (e.g., college students), with convergent validation in successive populations, until the measure is functioning at an acceptable level (see also Rogers, Harrell, & Liff, 1993).

References

Beaber, J. R., Marston, A., Michelli, J., & Mills, M. J. (1985). A brief test of measuring malingering in schizophrenic individuals. *American Journal of Psychiatry, 142,* 1478–1481.

Bernard, L. C. (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, L. C. (1991). The detection of faked deficits on the Rey Auditory Verbal Learning Test: The effect of serial position. *Archives of Clinical Neuropsychology, 5,* 81–88.

Bernard, L. C., & Fowler, W. (1990). 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.

Binder, L. M. (1990). Malingering following minor head trauma. *The Clinical Neuropsychologist, 4,* 25–36.

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

Binder, L. M., Villanueva, M. R., Howieson, D., & Moore, R. T. (1993). The Rey AVLT recognition task measures motivational impairment after mild head trauma. *Archives of Clinical Neuropsychology, 8,* 137–147.

Binder, L. M., & Willis, S. C. (1991). Assessment of motivation after financially compensable minor head trauma. *Psychological Assessment, 3,* 175–181.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (1982). *Test of Nonverbal Intelligence: A language-free measure of cognitive ability.* Austin, TX: Pro-Ed.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* San Diego, CA: Academic Press.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *The California Verbal Learning Test.* San Antonio, TX: Psychological Corporation.

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

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

Gillis, J. R., Rogers, R., & Bagby, R. M. (1991). Validity of the *M* Test: Simulation-design and natural-group approaches. *Journal of Personality Assessment, 57,* 130–140.

Goldberg, J. O., & Miller, H. R. (1986). Performance of psychiatric inpatients and intellectually deficient individuals on a task that assesses the validity of memory complaints. *Journal of Clinical Psychology, 42,* 792–795.

Hiscock, M., & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11,* 967–974.

Iverson, G. L., Franzen, M. D., & McCracken, L. M. (1991). Evaluation of an objective assessment technique for the detection of malingered memory deficits. *Law and Human Behavior, 15,* 667–676.

Lee, G. P., Loring, D. W., & Martin, R. C. (1992). Rey's 15–item visual memory test for the detection of malingering: Normative observations on patients with neurological disorders. *Psychological Assessment, 4,* 43–46.

Lezak, M. D. (1983). *Neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Pankratz, L. (1979). Symptom validity testing and symptom retraining procedures for the assessment and treatment of functional sensory deficits. *Journal of Consulting and Clinical Psychology, 47,* 409–410.

Pankratz, L., & Erickson, R. C. (1990). Two views of malingering. *The Clinical Neuropsychologist, 4,* 379–389.

Paul, D. S., Franzen, M. D., Cohen, S. H., & Fremouw, W. (in press). An investigation into the reliability and validity of two tests used in the detection of dissimulation. *International Journal of Clinical Neuropsychology.*

Rogers, R. (1988a). Introduction. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 1–9). New York: Guilford Press.

Rogers, R. (1988b). Researching dissimulation. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 309–327). New York: Guilford Press.

Rogers, R., Bagby, M., & Gillis, J. R. (1992). Improvements in the *M* Test as a screening measure for malingering. *Bulletin of the American Academy of Psychiatry and the Law, 20,* 101–104.

Rogers, R., Harrell, E. H., & Liff, C. D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review, 13,* 255–274.

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.

Wilhelm, K. L., Franzen, M. D., & Grinvalds, V. M. (1991, November). *Do people given knowledge and offered money fake better?* Paper presented at the 11th Annual Conference of the National Academy of Neuropsychology, Dallas, TX.

Received April 22, 1993
Revision received July 30, 1993
Accepted August 2, 1993 ■