The Clinical Neuropsychologist
1998, Vol. 12, No. 2, pp. 193-205

1385-4046/98/1202-193$12.00

Minding Your ''*p*s and *q*s'' When Using Forced-Choice Recognition Tests*

Richard I. Frederick and Robert L. Denney
U.S. Medical Center for Federal Prisoners, Springfield, MO

ABSTRACT

Denney (1996), and Frederick, Carter, and Powel (1995) described a forced-choice recognition test (FCRT) to evaluate suspicious complaints of remote memory loss. Although the two-alternative forced-choice trials in symptom validity testing have equal prior probabilities of correct answers (p) and incorrect answers (q), ps and qs might vary from 0.5 on individual trials of FCRTs. FCRTs nonetheless remain conservative procedures for evaluating suspicious memory loss, as long as the *overall p* approximates 0.5. In computer simulations, distributions in which $p \neq q \neq 0.5$ resulted in more conservative decision making than distributions in which $p = q = 0.5$. The authors demonstrate the ease in constructing FCRTs with overall $p = 0.5$ and discuss the proper application of FCRT findings in a clinical evaluation.

Denney (1996), and Frederick, Carter, and Powel (1995) described forced-choice recognition tests (FCRTs) that evaluate whether memories actually exist for suspicious complaints of amnesia. In general, the forced-choice procedure was developed to evaluate sensory stimuli (Haughton, Lewsley, Wilson, & Williams, 1979; Pankratz, Fausti, & Peed, 1975; Theodor & Mandelcorn, 1973) and is generally referred to as symptom validity testing (SVT; Pankratz, 1979). For instance, a questionable presentation of anesthesia can be evaluated by SVT through the presentation of 100 two-alternative forced-choice tactile discrimination tasks. On 50 trials, a tactile stimulus is presented. On the remaining 50 trials, no tactile stimulation is presented. The order of each type of trial is randomly assigned to preclude anticipatory responding. Individuals who truly have no tactile sense are limited to pure guessing (i.e., $p = 0.5$) and are expected to earn a score of 50 (i.e., Np).

In fact, the range of expected scores for persons with no tactile sense includes many scores in a range around 50. A two-tailed test of significance at alpha = 0.05 results in a range of expected scores of 42 to 58. Deviations outside this range allow reasonably firm conclusions that ability is present. Greater certainty can be obtained by expanding the range to limit the potential of random variation (e.g., at alpha = 0.01, the range is 40 to 60; at alpha = .005, the range is 37 to 63, both two-tailed tests). The probability that any score represents ''no ability'' can be evaluated by converting the earned score, number of trials, and probabilities of right or wrong answers to z scores by a formula given by Siegel (1956):

$$z = \frac{(x +/- .5) - Np}{\sqrt{Npq}}, \quad (1)$$

* We thank Robyn Dawes for reviewing an earlier version of this manuscript and for his helpful comments. This paper does not necessarily represent the views or official policies of the U.S. Department of Justice or the Federal Bureau of Prisons.
Address correspondence to Richard I. Frederick, Department of Psychology, U.S. Medical Center for Federal Prisoners, Springfield, MO 65807, USA. E-mail: rfrederi@ipa.net.
Accepted for publication November 9, 1997.

where N is the number of items administered, x is the obtained score, p is the probability of a correct answer, q is the probability of an incorrect answer ($p + q = 1$), and a correction of 0.5 is made to the obtained score to correct for the lack of continuity in scoring (add 0.5 to the score if $x < Np$, subtract 0.5 if $x > Np$). In two-alternative SVT, p and q are typically assumed to be equal to each other (i.e., $p = q = 0.5$), although SVT tasks can include any number of potential alternatives, as long as p and q are correctly computed. For example, a three-alternative forced-choice task results in $p = .33$ and $q = .67$.

Although Denney (1996) and Frederick et al. (1995) claimed that the FCRT could be used in the same manner as SVT, that claim has been sharply criticized by Dawes (1995). According to Dawes, in true two-alternative SVT tasks, p is always equal to q for naive subjects. The equivalence of p and q is based upon a random presentation of target items. For example, the Portland Digit Recognition Test (PDRT; Binder, 1993) is a SVT task. In the PDRT, test takers receive a stimulus item (a 5-digit number) and then must decide which of either a target or distractor is the same as the stimulus item. The placement of the target item (e.g., as potential answer ''top'' or ''bottom'') is decided randomly. Test takers with no true capacity to solve the test items have no basis on which to obtain a clue regarding the correct answer.

Dawes's criticism of the FCRT has to do with the potential answers for each trial. For example, on the FCRT, an examinee might be asked, ''What are you charged with doing?'' If the examinee claims not to know, he or she is presented with two alternatives (e.g., ''You are charged with distribution of cocaine or you are charged with distribution of methamphetamine. Guess which is correct.'') According to Dawes, there exists a potential bias in this form of questioning, because one answer might be more plausible, based on the nature of the world or based on the individual's own experiences. The following question highlights this criticism: ''Yes or no: Do the police report you were wearing a dress during the bank robbery?'' In this example, ''yes'' has an extremely low probability of being correct and a truly naive examinee who was attempting to obtain the best possible score would be compelled to answer ''no,'' which is the more plausible option.

We first note that both Frederick et al. (1995) and Denney (1996) cautioned against the use of questions with implausible alternatives, such as ''Do the police report you were wearing a dress during the bank robbery?'' Although we know of a case in which a bank robbery was committed by a man dressed as a woman, this question is otherwise heavily biased toward an answer of ''no.'' Secondly, we note that in our experience, it has been difficult to find alternative answers that seem as plausible as the correct answers. We have had to eliminate potential questions because the correct answer was obvious to the test taker, given the substance of information imbedded in the question. For example, a defendant was asked the color of the hat he was reported to have worn in a robbery. He said he did not know and he was then presented with two alternatives: ''red'' or ''blue.'' His response was to say, ''It couldn't have been blue. I never wear blue.'' Furthermore, Denney (1996) has directly tested the plausibility of answer alternatives he devised for three evaluations of suspicious amnesia complaints by administering the same questions to 60 adults with no prior knowledge of the case. The overall probability of a correct answer for all of his items (overall p) was about 0.5, ranging from 0.52 to 0.55 for the three evaluations.

In this paper, we will demonstrate the implications of p-value variations for items when p is systematically less than 0.5 or when overall p is equal to 0.5. (It should be obvious that when p is systematically greater than 0.5, the FCRTs will be extremely conservative and unlikely to detect feigned amnesia.) Following this, we will present a replication of Denney's (1996) findings regarding individual and overall p values for actual questions used in forensic evaluations. We will offer suggestions concerning the prudent development of potential questions for evaluation. Finally, we will describe how to evaluate FCRT results in the context of an evaluation.

When *p* is Systematically Less than 0.5

To conduct FCRT, one assumes that the overall *p* is equal to 0.5. When *p* is systematically below 0.5 for test items, but the *z* score is calculated as though the overall $p = 0.5$, the *z*-score values become inflated and falsely suggest improbable responding. For example, consider a situation in which the overall *p* for a set of 100 questions is actually 0.3, but the overall *p* is assumed to be 0.5. The questions are administered and 30 are correctly answered. How likely is the score of 30, given ''no memory''? If the *z* score is calculated with accurate knowledge of *p* and *q*, the *z* score is equal to 0, indicating that the performance is expected, given ''no ability'':

$$z = \frac{30 - 30}{\sqrt{100 * .3 * .7}} = 0.$$

If, however, the *z* score is computed under the incorrect assumption that $p = q = .5$ (i.e., the expected score is 50), then the earned score appears to be extremely unlikely:

$$z = \frac{30.5 - 50}{\sqrt{100 * .5 * .5}} = -3.9,\ p < 0.0000.$$

Thus, a systematic bias in *p* less than 0.5 can lead to frighteningly distorted conclusions about the presence of ability, and lead to damning conclusions regarding the effort of the examinee.

When Overall *p* = 0.5, but All Item Probabilities Are > 0.5 or < .05

We are not aware of any particular reason why *p* for an entire set of questions should vary substantially above or below 0.5 when thoughtfulness has been applied in the development of a question set. In other words, a *systematic* bias in *p* is never expected for prudent construction of a set of test items. *Nonsystematic* fluctuations in individual item probabilities are expected to occur, but the overall *p* value should remain at or about 0.5 (Denney, 1996). In the next section, we will evaluate the consequences of such nonsystematic and sometimes substantial variations among all test items probabilities for which the *overall p nonetheless remains equal to 0.5*. The examples we shall use will be extreme and unlikely to occur. We will demonstrate that the more extreme the deviation of *individual* items from $p = 0.5$ (either above or below this point), the more conservative decision-making will actually become.

The spread of any distribution of scores for a set of questions is a function of the standard deviation. As noted earlier, the expected standard deviation is derived from the overall probability of a correct answer, 0.5. But, when $p \neq q$ for any item, the expected standard deviation is given by first computing the sum of the products of individual item *p*s and *q*s:

$$\text{expected } sd = \sqrt{Npq} = \sqrt{\sum_{i=1}^{n} p_i q_i}, \tag{2}$$

the expected test mean is given by the sum of the individual item probabilities:

$$\text{expected test mean} = Np = \sum_{i=1}^{n} p_i, \tag{3}$$

and the correct *z* score is derived by substituting appropriately in equation (1):

$$z = \frac{(X +/- 0.5) - \sum_{i=1}^{n} p_i}{\sqrt{\sum_{i=1}^{n} p_i q_i}}. \tag{4}$$

For a 100-item test in which 20 questions each had a *p* value of 0.3, 0.4, 0.5, 0.6, and 0.7, the expected standard deviation would be:

$$\sqrt{20(.3)(.7) + 20(.4)(.6) + 20(.5)(.5) + 20(.6)(.4) + 20(.7)(.3)} = 4.8,$$

which is smaller than that expected for a 100-item test in which each item had a *p* value of 0.5 (expected $SD = 5.0$). Whenever individual *p*s $\neq$ 0.5 for any number of items, the standard deviation will always be smaller than when all *p*s = 0.5. In other words, when $p_1 \neq p_2 \neq p_i \neq p \neq q \neq$

0.5 for a FCRT, the distribution of scores for random responding should have a smaller range of scores because the standard deviation should be smaller than when $p = 0.5$. Because the range will be smaller, the likelihood of extreme scores being generated from random responding will be lower.

Equations 1 and 4 demonstrate that *z* scores are inversely proportional to the standard deviation. That is, as standard deviations become larger, *z* scores become smaller. So, when obtained scores are evaluated under the assumption that $p_1 = p_2 = p_i = p = q = 0.5$, no matter what the true situation is regarding p_1, p_2, or p_i, the *z* score is computed with the largest possible standard deviation and thus results in the smallest possible value of *z*, which means decision making is at its most conservative. To demonstrate that the distributions for scores obtained when $p \neq 0.5$ have a smaller range than when $p = 0.5$, we report a computer simulation in which both types of distributions were generated.

COMPUTER SIMULATION OF FCRTS

METHOD

Our goal was to generate two distributions in which the *overall p* value for a set of 100 items was equal to 0.5, but in which individual item probabilities were not equal to 0.5. For the first distribution (D_{37}) we wanted $p = .3$ for half the items and $p = .7$ for the other half of the items. For the second distribution (D_{46}) we wanted $p = .4$ for half the items and $p = .6$ for the other half of the items. We wanted to compare these two distributions to a third distribution (D_{55}) in which all *p* values, individual and overall, were equal to 0.5.

To generate all three distributions, we created five data sets of 50 items each. The data sets were intended to represent FCRT items of different *p* values. For example, for data set 1, we set *p* at 0.3 by giving 15 items a value of ''1'' (i.e., representing a correct response) and 35 items a value of ''0'' (i.e., representing an incorrect response). We set the remaining *p* values at 0.4, 0.5, 0.6, and 0.7 by forcing data set frequencies of ''1'' and ''0'' to correspond to those proportions. So, for example, data set 4 had a *p* value of 0.6, with 30 items equal to ''1'' and 20 items equal to ''0.''

Partial scores were derived by randomly sampling one item from a data set, observing its value, replacing the item in the data set, and resampling. After 50 trials, the values of each observation were summed to derive the partial score (minimum possible partial score, 0; maximum possible partial score, 50; expected value, *Np*). *Total scores* (minimum possible total score, 0; maximum possible total score, 100; expected total score, 50) were computed by combining two partial scores from a data set and its complement (e.g., data sets 1 and 5). To effectively approximate the normal curve for each distribution, we obtained a large number of total scores . To generate D_{37}, for example, we randomly matched 1000 partial score from data set 1 with 1,000 partial scores from data set 5. To create D_{55} (i.e., all *p*s = 0.5) we generated 2,000 partial scores from data set 3 ($p = 0.5$) and randomly combined them to make 1,000 total scores. We repeated this process 100 times, until we had 100,000 total scores for each distribution. Finally, each distribution (D_{37}, D_{46}, and D_{55}) was plotted to compare their dispersion of total scores.

RESULTS

The expected means of these total scores for each distribution was 50.0 (see Equation 3). The observed means were 49.8 for D_{37}, 49.9 for D_{46}, and 50.0 for D_{55}. The expected standard deviation for each distribution was 4.6 for D_{37}, 4.9 for D_{46}, and 5.0 for D_{55} (see Equation 2). The observed standard deviations, respectively, were 4.8, 4.9, and 5.3. The obtained dispersions of random scores are shown in the left panel of Figure 1. The center and right panels of Figure 1 highlight the differences in dispersion. As predicted, the distributions for D_{37} and D_{46} have a smaller range of total scores than D_{55}.

It is evident from Figure 1 that D_{55} has many more scores in the extreme range of below random scoring than D_{37} or D_{46}. We wanted to know the number of errors we would make for each distribution if we identified certain low scores as ''malingered.'' At the alpha = .05 level for this one-tailed test, the cutoff score (*x*) is 41 (41 or below representing an error), when equation 1 is solved for $z = -1.65$, $N = 100$, and $p = q = 0.5$. Following are the number of scores

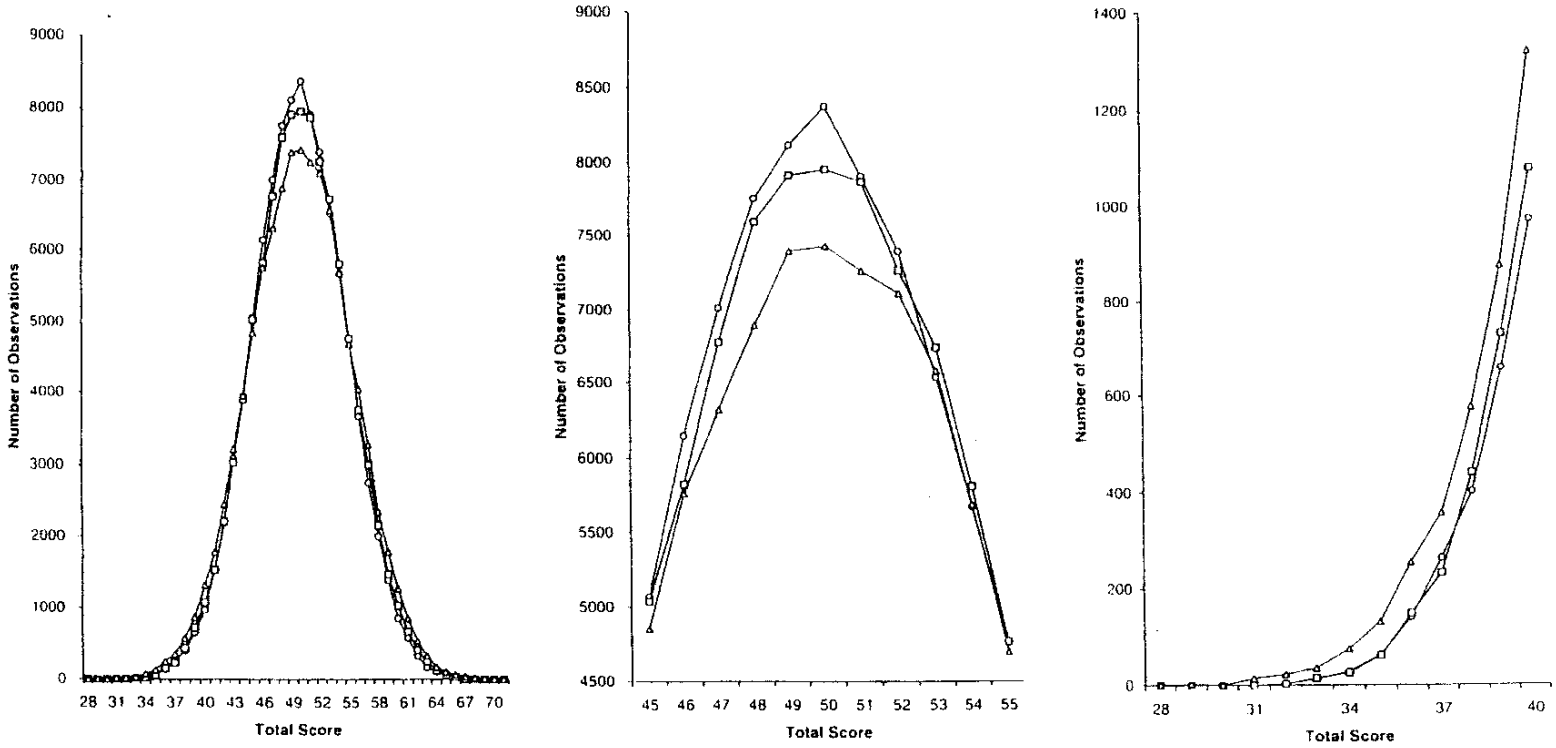

Fig. 1. Distributions of scores simulating random responding based upon underlying characteristics of p and q. Although overall $p = 0.5$ for all distributions, for D_{37} and D_{46} total scores were obtained from data sets in which $p_i = q_i = 0.5$. For D_{37}, $p = .3$ for one-half the data set and $p = .7$ for the other half of the data set. For D_{46}, $p = .4$ or $p = .6$. For D_{55}, $p_i = q_i = 0.5$. Left panel shows the dispersion of 100,000 total scores for each distribution. Center and right panels highlight distinctions in dispersions, demonstrating that D_{37} and D_{46} have fewer extreme scores than D_{55} ($\bigcirc = D_{37}$; $\square = D_{46}$; $\triangle = D_{55}$).

at or below 41 (errors in classification) for each distribution: D_{37} = 4,076; D_{46} = 4,274; and D_{55} = 5,455. The number of errors at alpha = .01 (cut-off score = 37) were D_{37} = 514; D_{46} = 487; and D_{55} = 899. As individual variation in item probability increases, the likelihood that total scores will be considered improbable decreases, when tested under the assumption that $p = q = 0.5$.

DISCUSSION

Nonsystematic variations in p and q for FCRT items do not make it more likely that individuals will be misclassified as malingering when they are only responding randomly. In fact, nonsystematic variations in p and q lead to more conservative classifications than SVT.

Examining Overall *p* for FCRTs

The only potential problem, then, in using FCRT is when p is substantially systematically biased so that the overall probability of a correct answer is less than 0.5. Denney (1996) demonstrated that actual tests used clinically to evaluate the probability of malingering had p values approximating 0.5. Given his findings, we doubt that there is much risk in generating an FCRT in which p is systematically lower than 0.5. In the second part of this paper we attempt to replicate Denney (1996) by testing this assumption once again.

GENERATING FCRTS

METHOD

We present three case examples to illustrate the utility of generating FCRTs to evaluate suspicious complaints of amnesia. The FCRTs were then administered to naive (no knowledge of the FCRT content) volunteers to approximate the potential biases in individual and overall p values for the FCRT items. Our hypothesis was that we would find that although individual p values might substantially deviate from 0.5, the overall p values would not. The case examples are from criminal forensic evaluations conduced for the U.S. District Courts. Appendix A includes examples of items used in these three clinical cases.

Case Examples

Case 1

Case 1 was charged with bank robbery and referred for a competency to stand trial evaluation. He had a long history of mental illness; past diagnoses included schizophrenia, bipolar disorder, and schizoaffective disorder. During the evaluation, he demonstrated no first-rank signs, but he displayed blunted affect and poor thought organization, with thought blocking and distractibility. Although he was considered to evidence legitimate severe mental illness, there were indications he may have been augmenting symptoms of illness and exaggerating memory loss. For example, when asked by the evaluator, he claimed nearly total memory loss about his younger life and parents, but he was able to provide meaningful personal and family medical history in a separate interview with a physician's assistant. He also claimed no memory for the alleged offense (investigative materials; e.g., videotape from bank surveillance cameras, clearly revealed his involvement in the robbery). He did not appear to be simulating cognitive impairment. Intellectual screening placed his functioning in the average range and other validity indicators were negative.

A 31-item FCRT was developed using information from investigative materials. Reasonably plausible alternate incorrect answers were devised. As the defendant claimed no recollection of events for the alleged robbery, as well as no recollection for information contained in the indictment and investigative materials, he was administered the test and asked to answer based on his memory of the investigative materials. He reasoned out the answers to three items so these items were excluded from the final test (Frederick et al., 1995). The defendant correctly answered 10 of the remaining 28 items to achieve $z = -1.32$ ($p = > .05$, one-tailed). His performance was considered to fall within the random range, and, consequently, could not support a conclusion of suppressed memory ability.

Case 2

Case 2 was referred for both competency-to-stand-trial and criminal responsibility evaluations. He was charged with drug possession, firearms violation, child pornography, and child molestation. He had no history of mental health treatment, although he reported numerous head injuries with loss of consciousness. He said he was experiencing audi-

tory hallucinations. He reported a history of significant alcohol abuse. On admission, he complained of profound memory loss for most of his life, and specifically for the time period of the alleged criminal activity. Physical assessment revealed glucose intolerance, treated by diet. EEG and CT scan of the head results were normal. Neurological examination revealed mild peripheral neuropathy. Scores obtained from intelligence testing were in the Borderline Retarded range. He identified six items and two rows on the Rey-15 Item Test (Pankratz & Binder, 1997). Dot Counting Test (Lezak, 1995) group time was almost the same as ungrouped time with 21 dots miscounted. He correctly identified 24 of 36 items on the abbreviated Hiscock Forced Choice Procedure (Guilmette, Hart, Giuliano, & Leininger, 1994) and demonstrated a negative slope. MMPI-2 scale *F T* score was 120, *F-K* raw score was 27, and total *T* score difference of obvious/subtle subscales was 320. On the Structured Interview of Reported Symptoms (SIRS; Rogers, Bagby & Dickens, 1992), he achieved two scales in the definite feigning range and four scales in the probable feigning range. The defendant's test performance and inconsistencies of behavior overwhelmingly indicated suppression of cognitive ability and memory as well as simulation of psychosis.

To assess his alleged memory loss more directly, a 35-item FCRT was developed from investigative materials. He correctly answered two items by using reasoning alone, so those items were eliminated from the final test. Of the remaining 33 items, he correctly answered 7, which resulted in a *z* score of –3.14 ($p < .001$, one tailed). Test results supported the opinion he was feigning amnesia.

Case 3

Case 3 was referred for a competency-to-stand-trial evaluation. Intelligence screening results varied with scores falling in below average, borderline retarded, and mild mental retardation ranges. He recalled two items from the Rey Auditory Verbal Learning Test (Lezak, 1995) but was only able to recognize five and correctly reject four words. He correctly identified only 17 of 36 items from the abbreviated Hiscock procedure. He achieved an MMPI-2 *F* scale *T* score of 120 with *F-K* raw score of 26. Test results suggested less than optimum effort in cognitive areas as well as simulation of psychosis. He appeared to be feigning remote memory loss as well.

A 28-item FCRT was developed from investigative material. He correctly answered only 4 items which corresponded to a *z* score of –3.59 ($p < 0.0001$, one-tailed). Results supported the opinion he was suppressing memory ability.

Participants

The three FCRTs described above were administered to a sample of 125 undergraduate and graduate students. Volunteers received nominal extra course credit. The sample consisted of 76 women and 49 men; 113 were White, 2 were African American, 3 were Asian American, 3 were Native American, 1 was Hispanic American, and 2 did not provide information about ethnicity. Mean age was 29 years ($SD = 10.9$). Mean years of education was 15.8 ($SD = 2.6$).

Procedure

Participants were told these sets of questions related to charges against criminal defendants. They were instructed that their involvement would assist to identify statistical properties of these tests. Participants were provided an answer sheet which did not include questions or answers; it only listed A and B for each item. They were instructed to provide their best guess for each question by circling A or B. The question was read to them, and the two possible answers were provided. After participants circled their answer, the correct answer was given. They were instructed to mark incorrect answers and total the number correct after each set of questions. They were then told that the next set of questions (the next FCRT) dealt with a totally new situation. This procedure more closely resembles the actual clinical setting than the procedure used previously (Denney, 1996) as participants did not have actual item answers printed on answer sheets. Four participants did not complete FCRT 1; one participant did not complete FCRT 3.

RESULTS

FCRT 1 contained 28 items; the overall *p* was 0.58. Individual item *p*s ranged from 0.21 to 0.93 (Table 1). FCRT 2 contained 33 items; the overall *p* value was 0.52. The range of individual *p*s was 0.26 to 0.74 (Table 1). FCRT 3 contained 28 items. The overall *p* value was 0.47; the range of individual *p*s was 0.12 to 0.75 (Table 1).

Table 1. Observed Rate of Correct Answers for FCRTs 1–3.

	Test 1		Test 2		Test 3	
Item	p^a	R	p^b	R	p^c	R
1	0.54	I	0.57	I	0.58	I
2	0.44	C	0.55	I	0.43	I
3	0.68	I	0.32	C	0.66	I
4	0.73	I	0.32	I	0.44	C
5	0.50	C	0.52	I	0.35	I
6	0.69	I	0.39	I	0.62	I
7	0.39	I	0.54	C	0.29	I
8	0.33	I	0.71	I	0.14	I
9	0.83	C	0.54	I	0.56	I
10	0.54	I	0.26	I	0.40	I
11	0.79	I	0.63	I	0.31	I
12	0.34	I	0.49	I	0.42	I
13	0.70	I	0.53	I	0.17	I
14	0.77	I	0.42	C	0.37	I
15	0.72	I	0.44	I	0.37	I
16	0.46	C	0.70	I	0.73	I
17	0.21	I	0.74	I	0.48	C
18	0.46	C	0.66	C	0.48	C
19	0.74	C	0.47	I	0.52	I
20	0.74	I	0.37	I	0.52	I
21	0.68	C	0.46	I	0.62	I
22	0.33	I	0.54	C	0.69	I
23	0.66	C	0.55	I	0.57	I
24	0.75	I	0.74	I	0.54	C
25	0.49	C	0.46	I	0.58	I
26	0.60	I	0.55	I	0.50	I
27	0.93	C	0.46	I	0.75	I
28	0.31	I	0.65	I	0.12	I
29			0.49	I		
30			0.73	I		
31			0.38	C		
32			0.39	C		

Note. p = observed incidence of correct answer for each item; R = nature of response by examinee: C = correct, I = incorrect; FCRT = Forced-Choice Recognition Test.
[a] $N = 121$.
[b] $N = 125$.
[c] $N = 124$.

DISCUSSION

These findings are consistent with the results of Denney (1996) and confirm our hypothesis that, although individual p values varied substantially from 0.5, the overall p values were not meaningfully different from 0.5. This should effectively eliminate the primary concerns of Dawes (1995). Not only is it relatively easy to construct FCRTs with overall p values of 0.5, there is no risk in false positive identification as a malingerer *because* of the variation in individual p values among the FCRT items. We have shown that unless the overall p for FCRT

questions is systematically biased, so that *p* is significantly less than 0.5, then the standard formula for *z* computation will provide an accurate (if not conservative) assessment of this likelihood.

We suggest several ways to prevent systematic bias in overall *p* for an FCRT. We believe the most important way to preclude systematic bias is to have an awareness of the potential for bias when preparing questions. Examiners should prepare questions so that the alternative to the true answer is readily *plausible*. For instance, suppose a defendant is charged with using a pistol to rob a bank. A *plausible*, but not necessarily equivalent, alternative to the true answer might be ''shotgun,'' or even ''machine gun,'' but ''knife'' does not seem as plausible. Any question in which an alternative is clearly more plausible than the correct answer should be eliminated from an FCRT. But we do not believe it is necessary to routinely subject proposed questions to a large sample of experimental participants to estimate ''true'' item probabilities (Cercy, Schretlen, & Brandt, 1997). We do believe, however, that it is prudent to have one or more colleagues review the questions, even to the point of administering the test to them out loud. Having colleagues defend their answers can often highlight problems in item construction. Finally, we suggest that the list of questions be as long as possible. This will ameliorate the effect of a small number of items which might be biased toward incorrect alternatives. Twenty five questions should be a reasonable minimum, because at that point the distribution of correct responses can readily be compared to a normal distribution (Siegel, 1956). We have used lists as long as 115 items. We note that making FCRT lists as long as possible greatly increases the likelihood that FCRTs will have overall *p* values greater than or equal to 0.5, because the *more plausible* answers are typically represented by the correct answers. Furthermore, a long list of questions ameliorates the potential effect of incorrect alternatives that are perceived as more plausible by naive test takers (e.g., individuals with true amnesia).

One cannot avoid all the internal biases or assumptions that examinees bring to an evaluation. Examinees, particularly those who are attempting to obtain a good score, may often anticipate whether an answer will be ''A'' or ''B'' based on a previous series of answers. For example, after the last three items in a series were correctly solved by choosing ''A,'' an examinee may be ''certain'' that the next answer must be ''B,'' thereby creating a bias to choose ''B.'' That bias does nothing to affect the a priori probability of correct answer assignment to ''B.'' (Similarly, a short trip to a roulette table will make clear that the intensity of any belief regarding the outcome of a spin of the wheel is completely independent of the prior probability of the outcome.) Consequently, we are relatively unconcerned about the variation in measured individual item probabilities for tests 1-3. We suspect that most deviations from 0.5 reflect a large contribution from sampling error, and that polling an infinite number of participants would lead to individual and overall probabilities indistinguishable from 0.5.

In conclusion, we encourage clinicians to evaluate the positive predictive power (PPP) and negative predictive power (NPP) of any test score when used to classify respondents. For the purposes of this paper, PPP refers to the probability that a positive score correctly identifies feigned amnesia; NPP refers to the probability that a negative score correctly identifies true amnesia or no impairment. PPP and NPP are not unchanging values (unlike sensitivity and specificity) and must be calculated based upon the local prevalence of the condition of concern and the prevalence of positive test scores in a local sample (Baldesserini, Finkelstein, & Arona, 1983). In our clinical setting, we can reliably estimate some of the prevalences necessary to compute PPP and NPP for positive FCRT scores:

(a) *Prevalence of true amnesia for events related to the instant offense*. We have had only a few individuals in the past several years who have demonstrated what appeared to be genuine partial amnesias for criminal conduct. The primary nature of amnesia for these individuals

Table 2. Observed Incidence of FCRT Scores for a Hypothesized Sample of 10,000 Criminal Forensic Evaluees.

Test result	With Amnesia	Without Amnesia	Total
Positive score	3	125	128
Negative score	97	9775	9872
Total	100	9900	10000

Note. See text. Prevalences of amnesia and scores are estimated based on our clinical experience. Prevalence of feigning estimated at 5%; probability of positive score if feigning estimated at 25%.

was blackouts for periods of intense intoxication. We liberally estimate the prevalence of true amnesia in our setting at 1%.

(b) *Prevalence of feigned amnesia for events related to the instant offense*. A review of completed clinical reports from 1990 to 1995 revealed 108 (12.1%) out of 893 individuals were labeled as malingering. Forty-three individuals (4.8%) were described as feigning amnesia or related memory disorders (in isolation or in addition to feigning cognitive impairment, psychosis, or multiple personality disorder). Consequently, we will estimate the prevalence of feigned amnesia in our setting to be 5%.

(c) *Prevalence of below random responding for individuals with true amnesia*. ''Below'' random responding is equal to the probability of ''above'' random responding. At overall alpha = .05, the probability of below random responding is 2.5%.

(d) *Prevalence of below random responding for individuals feigning amnesia*. SVT is considered a highly specific technique, but its sensitivity is considered rather low (Rogers, Harrell, & Liff, 1993). SVT may have higher sensitivity when the technique directly evaluates the feigned impairment. For example, Haughton et al. (1979) found that 55% of 20 individuals simulating hearing impairment were detected with audiometric SVT. Indirect evaluation of memory complaints with forced-choice memory tests has shown lower sensitivity. Binder (1993) found that 17% of mild head-trauma patients seeking compensation scored below chance on the PDRT. Guilmette et al. (1994) reported that 30% of 20 simulators scored below chance on the Abbreviated Hiscock test. Slick, Hopp, Strauss, Hunter, and Pinch (1994) observed that only 20% of feigners scored below chance on a revision of the Hiscock test. The FCRT directly evaluates a memory complaint for examinees who are likely much more highly motivated than simulators to prove impairment. Consequently, we believe the sensitivity of FCRTs is probably higher than for indirect SVT measures. Nevertheless, for the purposes of this article, we will conservatively estimate the prevalence of below random responding on the FCRT for feigning to be 25%.

Table 3. Observed Incidence of FCRT Scores for a Hypothesized Sample of 10,000 Criminal Forensic Evaluees.

Test result	Feigning	Not Feigning	Total
Positive score	125	3	128
Negative score	375	9497	9872
Total	500	9500	10000

Note. See text. Prevalences of feigning and scores are estimated based on our clinical experience. Prevalence of true amnesia estimated at 1%; probability of obtaining a positive score with true amnesia estimated at 2.5%.

(e) *Prevalence of below random responding for nonimpaired compliant test takers*. We consider the prevalence to be nonexistent, 0%.

Given these estimates, we can compute the PPP and NPP for both amnesia and malingering. Table 2 presents the prevalence of positive and negative test scores based on whether individuals do or do not have amnesia. A positive (below chance) score has a PPP for amnesia of only 2.3% (3/130; 2.3% of individuals with a positive score have amnesia) and a negative score (random or better than chance) has a NPP of 99.0% (9,775/9,872; 99% of individuals with a negative score do not have amnesia). Table 3 presents the same information based on whether individuals are or are not feigning amnesia. A positive score has a PPP of 97.7% for malingering; a negative score has a NPP of 96.2% (96.2% of individuals with a negative score are not feigning amnesia).

We share Dawes's concern against falsely accusing any individual of malingering (see also Steadman, 1980), but we believe the FCRT is inherently conservative, especially in the forensic evaluation context in which the rate of malingering amnesia is many times the rate of true amnesia. Given the rarity of true amnesia in our setting, a below-chance performance on the FCRT almost certainly results from individuals who can access memories. Certainty is increased whenever there is no satisfactory explanation for the genesis of the claim of amnesia. As noted in Frederick et al. (1995), however, the conclusion that such individuals are malingering must be resolved by an evaluation of the context in which the low score is obtained.

REFERENCES

Baldesserini, R. J., Finkelstein, S., & Arona, G. W. (1983). The predictive power of diagnostic tests and the effects of prevalence of illness. *Archives of General Psychiatry, 40*, 569-573.

Binder, L. M. (1993). Assessment of malingering after mild head trauma with the Portland Digit Recognition Test. *Journal of Clinical and Experimental Neuropsychology, 15*, 170-183.

Cercy, S. P., Schretlen, D. J., & Brandt, J. (1997). Simulated amnesia and the pseudo-memory phenomena. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 85-107). New York: Guilford Press.

Dawes, R. M. (1995, August). Final comments. In R. M. Dawes (Chair), *Courtroom testimony: Does our science justify continuing what we do?* Symposium conducted at the 103rd Annual Meeting of the American Psychological Association, New York.

Denney, R. L. (1996). Symptom validity testing of remote memory in a criminal forensic setting. *Archives of Clinical Neuropsychology, 11*, 589-603.

Frederick, R. I., Carter, M., & Powel, J. (1995). Adapting symptom validity testing to evaluate suspicious complaints of amnesia in medicolegal evaluations. *Bulletin of the American Academy of Psychiatry and the Law, 23*, 231-237.

Guilmette, T. J., Hart, K. J., Giuliano, A. J., & Leininger, B. E. (1994). Detecting simulated memory impairment: Comparison of the Rey 15-Item Test and the Hiscock Forced-Choice Procedure. *The Clinical Neuropsychologist, 8*, 283-294.

Haughton, P. M., Lewsley, A., Wilson, M., & Williams, R. G. (1979). A forced-choice procedure to detect feigned or exaggerated hearing loss. *British Journal of Audiology, 13*, 135-138.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Pankratz, L. (1979). Procedures for the assessment and treatment of functional sensory deficits. *Journal of Consulting and Clinical Psychology, 47*, 409-410.

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

Pankratz, L., Fausti, S., & Peed, S. (1975). A forced-choice technique to evaluate deafness in the hysterical or malingering patient. *Journal of Consulting and Clinical Psychology, 43*, 421-422.

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

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.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Slick, D., Hopp, G., Strauss, E., Hunter, M., & Pinch, D. (1994). Detecting dissimulation: Profiles of simulated malingerers, traumatic brain-injury patients, and normal controls on a revised version of Hiscock and Hiscock's forced-choice memory test. *Journal of Clinical and Experimental Neuropsychology, 16*, 472-481.

Steadman, H. J. (1980). The right not to be a false positive: Problems in the application of the dangerousness standard. *Psychiatric Quarterly, 52*, 84-99.

Theodor, L. H., & Mandelcorn, M. S. (1973). Hysterical blindness: A case report and study using a modern psychophysical technique. *Journal of Abnormal Psychology, 82*, 552-553.

APPENDIX A

This appendix contains sample items used in the FCRT for Case 1. The correct alternatives are marked with asterisks.

1. The robbery allegedly occurred on what date?
   a. May 19, 1995*
   b. April 30, 1995
2. The robbery allegedly occurred at about what time?
   a. 10:30 am
   b. 4:45 pm*
3. Initially, the robber was alleged to do what?
   a. Sit in a chair
   b. Stand at the deposit table*
4. The robber then did what?
   a. Spoke to the teller*
   b. Handed the teller a note
5. Was the teller a man or woman?
   a. Man
   b. Woman*
6. Was the robber allegedly wearing a hat?
   a. No
   b. Yes*
7. What type of hat was he wearing?
   a. Sock cap
   b. Baseball-style cap*
8. Was he wearing a coat?
   a. No
   b. Yes*
9. What color was the coat?
   a. Brown*
   b. Grey
10. How much money did the robber allegedly get?
    a. $1,183*
    b. $2,670