# On the Interpretation of Below-Chance Responding in Forced-Choice Tests

**Richard I. Frederick**
*U.S. Medical Center for Federal Prisoners*

**F. Michael Speed**
*Texas A&M University*

*Two-alternative, forced-choice tests are commonly used to assess cooperation in examinations of neurocognitive functioning. Most commercially available tests do not primarily depend on comparing the total correct responses to the number expected by guessing. Nevertheless, the tests afford an opportunity to make stronger judgments about the cooperation of test-takers when the test score is lower than the range of scores expected for guessing. Unfortunately, many researchers and clinicians make serious errors in communicating what is "guessing" and what is "worse than guessing" (or malingering). This article describes proper methods of evaluating total correct responses on a forced-choice test.*

*Keywords:* forced-choice testing; malingering; neuropsychology; hypothesis testing; psychological evidence

This article concerns the detection of feigned cognitive impairment using two-alternative, forced-choice testing and the limits of interpretations for performances on these tests. Feigned cognitive impairment, or malingering, will be referred to as an intention to underrepresent one's true maximal capacity to answer correctly on tests of cognitive capacity (Frederick, 2003). The relatively high prevalence of malingering in diverse forensic clinical examinations (e.g., see Mittenberg, Patton, Canyock, & Condit, 2002) and the limitations of subjective clinical judgment produce a need for objective validity indicators in the assessment of cognitive functioning (Bigler, 1990; Faust & Guilmette, 1990). Pankratz, Fausti, and Peed (1975) introduced two-alternative, forced-choice testing (FCT) as a methodology to assess feigned presentation of psychophysical complaint. Building on the work of Brady and Lind (1961), Grosz and Zimmerman (1965), and Theodor and Mandelcorn (1973), Pankratz (1979) developed FCT as an objective method of evaluating suspicious sensory complaints by examining response patterns on FCT sensory discrimination tasks.

FCT involves the presentation of a series a trials in which patients are asked to choose between two alternatives (i.e., a correct response—the target and an incorrect response—the distractor or foil). The inherent difficulty of the trial is relatively low, so low that even individuals with limited ranges of ability can identify the correct response (Faust & Auckley, 1998). For example, Slick et al. (2003) demonstrated that even individuals with severe memory impairment achieved perfect or near-perfect scores on a forced-choice digit-recognition test.

The trials within an FCT are construed to have equivalent difficulty. The number of correct responses is typically the score of interest. Scores lower than expected for guessing (i.e., below-chance responding) are considered specific to malingering. Because of the extremely low or absent false positive rate associated with conclusions that below-chance scores represent malingering, below-chance responding is promulgated as a basis in the Slick criteria for malingering (Slick, Sherman, & Iverson, 1999) to determine that malingering is probable or definite, depending on other factors. In early FCT procedures, persons being examined were told trial-by-trial whether they had responded correctly or not to induce a sense of doing "too well" among those who intended to do poorly. This practice remains effective and recommended when possible (e.g., Frederick, Carter, & Powel, 1995; Frederick & Denney, 1998; Tombaugh, 1997).

*Assessment*, Volume 14, No. 1, March 2007 3-11
DOI: 10.1177/1073191106292009

Key aspects of this methodology have been successfully adapted to the assessment of malingered memory deficits (e.g., Binder, 1993; Hiscock & Hiscock, 1989; Iverson, Franzen, & McCracken, 1994; Slick, Hopp, Strauss, Hunter, & Pinch, 1994; Tombaugh, 1997), malingered amnesia (Frederick & Denney, 1998; Frederick et al., 1995), and malingered cognitive deficits (Frederick, 1997, 2003; Frederick & Crosby, 2000; Frederick, Crosby, & Wynkoop, 2000). Currently, the application of FCT for the assessment of cooperation in forensic assessments is common, if not routine (Binder, 2002; Millis & Volinsky, 2001). Generally, the method of discriminating among genuine and incredible performance is by total correct responses, even when there is no below-chance responding. Empirically derived cutoff scores are well-established for the most common procedures (Binder, 2002; Etcoff & Kampfer, 1996; Frederick, 2003; Tombaugh, 2002). Some researchers restrict dissemination of cutoff scores to preclude potential fakers from using the information to more effectively malinger impairment (Binder, 2002) by access through the Internet or by other means (e.g., see Ruiz, Drake, Glass, Marcotte, & van Gorp, 2002; Tan, Slick, Strauss, & Hultsch, 2002).

Some researchers have adapted FCT procedures to induce a sense that trials have differing difficulties, comparing differences in response accuracy for easy and difficult trials (Binder, 1990; Thompson, 2002). Others have created trials to have a true hierarchy of difficulty across trial and then used distinctive features of average performance accuracy plotted against average trial difficulty to identify problematic response styles (Frederick & Crosby, 2000; Frederick et al., 2000).

Notwithstanding these potentially powerful modifications to FCT methodology, each FCT procedure retains the capacity to identify malingering on the occasion of the below-chance score. For example, many commercially available tests (e.g., the Portland Digit Recognition Test, PDRT; the Test of Memory Malingering, TOMM; and the Validity Indicator Profile, VIP) do not use below-chance responding as the primary detection strategy, but each test can make the stronger FCT argument in some circumstances. That is, persons who complete these tests sometimes answer fewer items correctly than would be expected for guessing. This article concerns issues related to the interpretation of such below-chance responding.

## CALCULATING THE PROBABILITY OF A RANDOM EVENT

FCT evaluates a test score (total correct responses) by comparing it to the number of correct responses expected for guessing. Random responding (also referred to as irrelevant responding or content nonresponsive responding) on an FCT produces what is commonly referred to as "chance" distribution of scores. That is, various combinations of right and wrong answers for random responding are binomially distributed, and this distribution approximates a normal distribution when a sufficient number of test items are sampled. The most commonly occurring number of correct responses, the mode, for true random responding on a FCT is half the number of trials, $Np$,[1] the number of trials (test items) on the test ($N$) multiplied by the probability of choosing a correct answer by guessing alone ($p$, usually $p = .5$). Individuals who guess on a two-alternative FCT are expected to generate an equal number of correct responses and incorrect responses. (The expected number of incorrect responses is given by $Nq$, where $q = 1 - p$.)

The assumption that any observed number of correct responses has been randomly generated allows for testing of that assumption by comparison to two particular distributions that model random events, the binomial distribution and the normal distribution. Methods for calculating the probability of observing a certain score or lower in reference to these distributions are given in the appendix. Alternatively, a search of the Internet will readily locate Web sites that use these methods to compute the $p$ value, which stands for the probability that the observed number (or lower) of correct responses were generated by guessing alone.

## WHAT IS BELOW-CHANCE RESPONDING?

Even though it is the most commonly observed individual score, $Np$ itself will occur only a small fraction of the time with respect to all other scores, both above and below $Np$. Chance level actually comprises a range of scores symmetrically distributed around $Np$. We will refer to this as the "random range." This range of scores reflects the desired alpha level and the expected standard deviation, which is given by $\sqrt{Npq}$. Scores in the lower half of this range (below $Np$) are much more likely to be observed than $Np$ itself. For example, consider an FCT procedure in which $N = 20$, $p = .5$, and $Np = 10$. Tables of binomial expansion (e.g., Hays, 1973, Appendix C, Table 2) report that the likelihood of a score of 10 on a FCT test of 20 items, assuming random responding, is about 18%. This means that about 82% of the time for this combination of $N$ and $p$, scores generated by random responding will be greater than or less than $Np$; 41% of the time such scores will be lower than $Np$. To repeat, $Np$ is the most commonly observed individual score, but we are much more likely to see scores above or below $Np$ than to actually observe $Np$. We are generally only interested in scores lower than $Np$ when they fall below the lower endpoint of the range expected for guessing. (An exception might be for repeated testing in which scores for the same

individual consistently fall below *Np* but do not fall below the lower endpoint of the random range.)

In general, there are only two influences that contribute to total scores falling below the lower endpoint of the random range. The first is random responding. We expect random responding itself to result in some scores falling below the lower endpoint of the random range at the rate of alpha or one half alpha, depending on whether our null hypothesis is directional or nondirectional. The second potential influence for below-chance responding is malingering—the willful choosing of incorrect responses.[2] We note that random responding itself often is considered to be a form of malingering (e.g., as in the Slick criteria) because most FCT trials generally have only a token level of difficulty. Consequently, a total lack of capacity to contend with the test content would reflect impairment so severe as to be obvious in casual interaction. In fact, the PDRT (Binder, 1990; Binder & Willis, 1991) and the TOMM (Tombaugh, 1997) use primary detection strategies that are predicated on the observation that even individuals with significant impairment routinely perform significantly better than chance. In forensic assessment, random responding generally reflects a refusal to engage in the task or a desire to appear impaired. That is, in most forensic clinical examination settings, we expect that the local base rate for individuals with no ability to contend with test items is generally greatly exceeded by the local base rate for individuals who refuse to engage in the task or who desire to appear impaired. For example, Frederick and Denney (1998) reviewed a sample of 893 consecutively evaluated criminal defendants and estimated that about 5% were feigning aspects of amnesia but far fewer than 1% actually had amnesia.

## WHAT ARE OUR HYPOTHESES?

When we generate probability values, we must be certain what hypotheses we are testing. In using FCT procedures, we generally are evaluating one of two beliefs. The first belief is that the person does not have the capacity to contend with the test items. For instance, let us assume that we are evaluating an individual who claims to have amnesia and we do not have any basis to strongly suspect deception. We decide to test this claim through the sort of FCT procedure developed by Frederick et al. (1995) and discussed in Denney (1996) and Frederick and Denney (1998). Our hypotheses are articulated as follows:

$$H_o: p = .5.$$

$$H_A: p \neq .5.$$

What we mean by $p = .5$ is that the probability of getting a correct response on a trial by guessing is .5. Given our assumption that $p = .5$, we are interested in the likelihood of observing *x* or fewer correct answers *or x* or more correct answers out of *N* trials when this null hypothesis is true.[3] Our very firm belief is that for each of the individual trials themselves, $p = .5$. That is, we believe that nothing about the trial itself induces an individual to pick one choice over another. Furthermore, we conclude, with error rate alpha, that departures from $p = .5$ reflect contributions from the individual being tested and not from some idiosyncrasy in the test item.[4] Because we believe that the trials always have $p = .5$, our tests of $p = .5$ are really tests about the individual's contribution to the number of correct responses. When $p > .5$, we believe the person has the capacity to answer correctly and chose to do so. When $p < .5$, we believe the person has the capacity to answer correctly and chose to answer incorrectly. To make our assumptions evident, we will use the notation $p_i$ to reflect that we are evaluating characteristics of the individual being tested, not the test items themselves. That is, $p_i$ is the probability that the individual being tested will choose the correct response, based on the individual's own ability and level of cooperation.

Continuing with the example of testing amnesia, when we believe an individual may actually have amnesia and possesses no ability to identify the correct responses on their personalized memory tests, we articulate our hypotheses in this way:

$$H_o: p_i = .5$$

$$H_A: p_i \neq .5$$

To test this null hypothesis, we are interested in deviations above or below the random range. If we are using the *z* statistic and an $\alpha = .05$, our boundaries for the *z* distribution are $\pm$ 1.96. Values of *z* must be more extreme than these to reject the null hypothesis.

The second belief we might have is that the individual is only claiming to have amnesia but actually has intact memories. We might have this idea based on some aspect of presentation in interview, some information we gained through a review of past history, or some suspicious performance on another test. We believe that the individual will attempt to support the claim of amnesia by generating only a relatively small number of correct answers on our FCT procedure, even though he or she could correctly respond to most, if not all, of the trials. Because we have no interest in scores greater than the upper limits expected for guessing, nondirectional hypotheses unnecessarily limit our capacity to detect malingered presentations. Directional hypotheses shift our alpha risk such

that our decision point has lower magnitude, making it more likely that we will reject the null hypothesis. We rewrite our hypotheses in this way:

$$H_o: p_i \geq .5$$

$$H_A: p_i < .5$$

To test the null hypothesis, we are interested only in deviation below the random range. If we are using the $z$ statistic and an $\alpha = .05$, our boundary for the $z$ distribution is –1.645. Values of $z$ must be lower than this to reject the null hypothesis.

## PROBLEMATIC INTERPRETATIONS OF BELOW-CHANCE RESPONDING

What constitutes below-chance responding has sometimes been portrayed inaccurately in the literature. For example, Meyers and Volbrecht (1998) identified below chance responding as "10 out of 20 or less" and considered such responding to be "suggestive of malingered performance" (p. 304), even to the point of potential utility as a "gold standard" of malingering (p. 306). This position has been supported by Reynolds (1998), who stated, "any score with a chance probability of less than .5 would merit at best judgment or classification as malingering" (pp. 273-274).

The position of Meyers and Volbrecht is clearly wrong. A circumstance in which $p_i \leq .5$ does not constitute below-chance responding. Below-chance responding is restricted to circumstances in which $p_i < .5$ at some acceptable alpha level. Reynolds may be making the same argument as Meyers and Volbrecht, in which case his assertion also would be incorrect. But Reynolds may mean that *malingering* is identified when $p_i \leq .5$. As noted earlier, it may be characteristic of malingerers to randomly respond (i.e., $p_i = .5$), but other clinical conditions also may produce $p_i = .5$ (e.g., mental retardation). Without speaking to this latter potential for misclassification, Reynolds may be testing a different set of null and alternative hypotheses:[5]

$$H_o: p_i > .5$$

$$H_A: p_i \leq .5.$$

Here, the null hypothesis asserts that only above-chance responding reflects cooperation; all other responding is uncooperative. This construction provides more opportunities to identify malingering, defined here as performances that are worse than chance as well as performances that cannot be differentiated from chance. In such a construction, in only those circumstances in which there is clearly better-than-chance responding will we conclude that the individual was cooperating. If we are uncomfortable with calling any FCT performance malingered unless the individual has clearly shown the capacity to answer correctly,[6] we should be more careful in how we define at chance or below. It is not correct to say that at chance or below translates into an observed $p_i$ of .5 or below. At chance or below might actually reflect observed $p_i$s of .6 or below, depending on the number of trials administered. For example, for 100 trials, the range of random responding at $\alpha = .05$ generates a range of 40 to 60 correct responses ($p_i = .4$ to .6); at chance or below is $p_i \leq .6$.

This sort of misunderstanding about what constitutes below-chance responding is by no means limited to researchers and clinicians. Binder (2002) reported that a judge concluded that a person was faking impairment because the PDRT score was 32 out of 72. The judge wrote, "It is obvious that someone should get at least half of the answers right. . . . Random chance should not result in a score of 32" (p. 37). Although Binder did not characterize this as obviously faulty reasoning, we note that the probability of generating a score of 32 or lower when $p_i = .5$ is .20. The probability of earning a score = $Np \pm 4$, or 32 to 40, when $p_i = .5$, is .71. That is, guessing will result in scores 32 to 40 about 71% of the time.

Reitan and Wolfson (1996) claimed some clinical conditions result in $p_i < .5$: "Clinical observations of some severely brain-damaged persons, however, indicate that significant neuropsychological impairment can also *cause* a subject to perform below chance levels" (p. 574, Italics added). The idea that pathology causes below-chance responding is tantamount to saying that certain clinical conditions result in an increased likelihood of choosing incorrect answers in circumstances in which the presence of no ability otherwise results in $p_i = .5$. Severe pathology can cause an individual to perform below chance only in the same way that it can cause an individual to perform above chance. When we hypothesize $p_i = .5$, for whatever reason, we do expect to observe below-chance performances at a rate consistent with our chosen alpha. For example, Frederick (1997, 2003) reported that 1 of 40 individuals (2.5%) with mental retardation (many of whom were illiterate) generated a score below chance on an FCT measure of vocabulary. We do expect to observe below-chance responding at a rate consistent with alpha, but we do not conclude that random responding causes the rate of correct choices to fall above or below the demarcated random range.

Most commonly, we believe that individual ability levels result in $p_i > .5$. As noted earlier, on the PDRT and the TOMM, research has established that even for significantly impaired individuals, $p_i$s are generally much

higher than .5. In scholastic settings, true-false tests are popular methods of assessing ability because when student $p_i$s > .5, tests scores will easily exceed values expected for random responding.

In fact, not only do we commonly assume that individual ability levels govern $p_i$, we recognize that most circumstances in which $p_i$ is outside the range expected for guessing ($p_i \pm 5$) involve instances in which ability is present, regardless of whether the observed $p_i$ is greater or less than .5. Consequently, for a 100-item test, a score of 100 has the same meaning as a score of 0 with respect to ability. Deviations of equivalent magnitude but opposite valence have the same meaning with respect to ability, although they most likely have opposite meanings with respect to intention to respond correctly (Frederick, 1997; Frederick et al., 2000). Thus, the second potential influence on $p_i$, in addition to ability, is the intention of the test-takers—whether they intend to respond correctly or do not intend to respond correctly. When test-takers demonstrate $p_i$s significantly greater than .5, we not only assume they had the ability to respond correctly but we assume they chose to do so.

The third potential influence on $p_i$ is the effort of the test-taker, the sustained intensity by which individuals apply their abilities and intention to the task at hand. When ability is present, and the test-taker has some intention to respond correctly, in most FCT applications, the value of $p_i$ is determined by effort, the intensity of application of ability in the intended direction (Frederick et al., 2000). Frederick (1997, 2003) has proposed that performances on two-alternative, forced-choice tests can be classified by a cross-classification of intention and effort (compliant, strong effort to respond correctly; inconsistent, weak effort to respond correctly; irrelevant, weak effort to respond incorrectly; and suppression, strong effort to respond incorrectly).

In summary, scores that are outside a range of score values demarcated around *Np* are a gold standard only of improbability with respect to the notion that no ability exists ($p_i = .5$). In most instances of FCT administrations, scores are above chance. We generally conclude in such instances that the results are so improbable for guessing that some ability to respond correctly exists, that there is at least some willingness to respond correctly, and that some level of effort has been applied to respond correctly. When scores are below chance, we generally conclude that the results are so improbable for guessing that some ability to respond correctly exists, that there is an intention to respond incorrectly, and that some level of effort has been applied to respond incorrectly. Confidence about these conclusions increases as the scores increasingly deviate from chance beyond our demarcated range.

## ASSUMPTIONS OF FCT FOR MALINGERING DETECTION

In summary of the above discussion, we identify three primary assumptions that underpin forced-choice testing in malingering detection.

*Assumption 1.* In the absence of any ability to correctly solve the test items, $p_i = 1/n$, where $n$ is the number of alternatives to choose among in any trial. For two-alternative FCT procedures, $n = 2$ and $p_i = .5$.

Our first assumption is that in the presence of no ability to correctly solve the FCT trial, both $p = .5$ and $p_i = .5$. In other words, we assume that there is nothing inherently biasing about the construction of the test item and that individuals who lack the capacity to solve the test items will guess at the answers. There are some circumstances in which the content area of FCT procedures present challenges to this assumption. Frederick and Denney (1998) discussed the challenges of creating FCT procedures to assess feigned amnesia so that each trial had a $p = .5$. Variations in $p$ occur because one answer might be more likely to be chosen than another based on social desirability or perceived prior probabilities. Frederick and Denney also demonstrated that nonsystematic variations in the probability of a correct response for individual items variation about $p = .5$, above or below, did not preclude standard hypothesis testing as long as the overall $p$ for all trials was equal to .5.[7]

*Assumption 2*. The total number of correctly solved items, when $p_i \neq .5$, represents a combination of items correctly solved by ability and items correctly solved by guessing.

Our second assumption is that the total score reflects a combination of the number of items answered correctly by knowledge and the number of items answered correctly by guessing. In general, this sum can be understood as follows:

Total Score = Number of items correctly solved
+ Number of items correctly guessed.

Frederick (2003) showed that in a two-alternative FCT procedure, a simple transformation allows us to estimate the number of items correctly solved:

Number of items correctly solved = Rights – Wrongs.

For example, in a test with 100 items, a score of 50 estimates that the individual could not solve any of the test

items (i.e., 50 – 50 = 0) and guessed at them all. A score of 75 suggests the individual could solve 50 items (i.e., 75 – 25 = 50). The individual is estimated to have solved 50 items and guessed at 50 items, getting 25 correct: 50 + 25 = 75. A score of 100 indicates perfect knowledge, but so does a score of 0. The score of 0 is just as difficult to obtain as the score of 100. In these instances, we conclude $p_i \neq .5$ (i.e., when the score = 100, we conclude $p_i = 1$; when the score = 0, we conclude $p_i = 0$). We cannot be exactly sure how scores are derived. Our best estimate remains as follows:

Number items correctly solved
+ Number of items correctly guessed.

*Assumption 3*. The total number of correctly solved items does not necessarily represent the best abilities of the individual being tested.

When the number of items correctly solved deviates from chance, $p_i \neq .5$, there is likely a contribution of ability to the total score. When the score exceeds chance (when $p_i > .5$), the intention of the test-taker is to correctly solve at least some of the items. When the score is below chance (when $p_i < .5$), the intention of the test-taker is to incorrectly solve at least some of the items.

As the number of correctly solved items deviates more extremely from chance, whether above or below chance, the contributions of ability and effort increasingly account for the extremity of deviation. That is, the extremity of deviation is not likely a consequence of changes in intention.

Even perfect performance on many malingering tests or effort tests does not routinely indicate that we have measured the maximal abilities of the individual or that maximal effort was even required to respond correctly (Faust & Auckley, 1998). The inherent difficulty of most of these tests is minimal. In fact, many malingering tests routinely expect performances of near 90% for compliant individuals, even if some impairment exists (Tombaugh, 2002).

## POST HOC INTERPRETATION

When one uses an FCT to investigate a specific hypothesis, the process of interpretation is straightforward. If one rejects the null hypothesis, one accepts the alternative hypothesis. But what is the proper method of interpretation when one is really not using an FCT procedure to test a null hypothesis but is using an FCT procedure as a sort of screening process? For example, several tests, the PDRT, the TOMM, and the VIP (Frederick, 2003), do not appeal to below-chance responding as a primary method for evaluating the validity of response on cognitive and memory testing. The TOMM and PDRT are considered positive (i.e., likely feigned) when the number of correct responses is above chance but below a particular cutoff score. The VIP uses primarily a performance-curve strategy to evaluate the validity of responses. These tests are often administered routinely in thorough evaluations, without a priori hypotheses that individuals are faking. Suppose we are using the primary detection strategies of the test, which do not depend on comparisons to chance, but we notice that a score is below chance. Are we able to make use of this additional information?

Obviously, we can compute the probability that the score would have occurred randomly, but what is our null hypothesis and what is our alternate hypothesis? Can we construct the null hypothesis that $p_i \geq .5$ and alternate hypothesis that $p_i < .5$ after we have seen the score? Will this give us an unfair advantage in setting our alpha risk (e.g., $\alpha = .05$) and the corresponding critical point (e.g., $z = -1.645$ vs. $z = -1.96$)? Our answer is, if our desired alternate hypothesis is $p_i < .5$ (this means we are interested only in below-chance responding), then our null hypothesis is $p_i \geq .5$.

$$H_o: p_i \geq .5.$$

$$H_A: p_i < .5.$$

We use this form of hypothesis testing when we have no a priori basis to suspect that performances are malingered. We commit to our same level of risk that we typically endorse (we suggest $\alpha = .05$ with critical point $z = -1.645$).

If, however, for other reasons we suspect malingering, our post hoc analyses of FCTs have a null hypothesis that the person is malingering and our alternate hypothesis is $p_i > .5$ (i.e., chance responding or worse represents malingering; better than chance responding represents cooperation). Then, our null hypothesis is $p_i \leq .5$.

$$H_o: p_i \leq .5$$

$$H_A: p_i > .5$$

Consequently, if for other reasons we suspect malingering and observe random or below-chance responding on an FCT, we have supported our suspicion. Such practice is quite sensible; some researchers consider random responding to represent a reliable indicator of malingering (Lees-Haley, Dunn, & Betz, 1999). We commit to our same level of risk that we typically endorse (we suggest $\alpha = .05$). We are interested in the likelihood of observing $x$ or fewer correct answers out of $N$ trials when this null hypothesis is true. What is $x$? It is that score at the boundary of chance

responding and better-than-chance responding; our critical point is $z = 1.645$.

Reynolds (1998) has suggested that the alpha level (i.e., $\alpha = .05$) need not be stringent in post hoc analyses, particularly when other evidence suggests malingering. For example, Reynolds stated that on a four-alternative test (i.e., $N = 40$, $p = .25$, $Np = 10$), the probability of earning a score of 7 out of 40 is sufficiently improbable (significance level for 7 or lower = .182), such that with other evidence of inconsistent performance, a conclusion of malingering is warranted. Binder (2002) also supports this notion, setting alpha to .10 ($p = .35$).

We disagree. First, if we have other evidence of malingering, the best thing we can do is form our hypotheses about a particular FCT prior to "peeking" at its test results. This mitigates problems associated with post hoc analysis. Second, there are many legitimate reasons to modify our chosen risk levels (change the values of alpha) in research, but increasing the standard value of alpha post hoc is rife with implications of data fitting. Relaxing alpha in this way means only that we are willing to consider scores closer to *Np* as potential indicators of malingering. But, as we earlier demonstrated in a discussion of the PDRT, scores just a little lower than *Np* (i.e., 32 out of 72) should be interpreted the same ways as scores just a little bit higher than *Np* (i.e., 40 out of 72). Consequently, this sort of inquiry is not a matter of relaxing alpha to increase beyond the standard .05; it is a matter of clearly stating our null hypothesis, $p_i \leq .5$. For example, for the 72 trials of the PDRT, it requires 44 or more correct responses to reject the null hypothesis. In the example generated by Reynolds, we would agree that 7 out of 40 was consistent with malingering, but we note that we would have failed to reject our null hypothesis, $p_i \leq .5$, only when the person correctly answered as many as 15 or more out of 40.

## COMMUNICATING THE RESULTS OF OTHERS TO DECISION MAKERS

In forensic assessments, decision makers include judges, juries, worker's compensation boards, disability panels, and insurance adjustors. Others not in a decision-making capacity, including other clinicians and attorneys, often review our work. How shall we communicate the findings of FCT assessments, particularly when we observe below-chance responding?

Whether we generate a probability value through the binomial expansion or by conversion from a computed *z* score, we must be certain what this probability value means. If the total score is 17/50 and the computed probability is .016, what does this mean? It means that if an individual does not attend to item content and merely picks an alternative at whim throughout the entire test, we will observe scores at 17 or lower less than 2% of the time. *It is an improbable event*. It does not mean that there is a 98% probability of faking. We should not conflate the sensitivity (the probability a malingerer will score below chance) of the FCT procedure (which is generally fairly low; Rogers, Harrell, & Liff, 1993) with its positive predictive power (the probability that a below-chance score represents malingering; see, e.g., Baldessarini, Finkelstein, & Arana, 1983, Elwood, 1993; Frederick, 2000; Rosenfeld, Sands, & van Gorp, 2000).

What we want to know in the clinic is the probability that scoring below the cutoff represents faking (i.e., positive predictive power). There is a basis to explain to others the probability that scoring below the cutoff represents faking, but the following information is required:

1. The prevalence of malingering in such evaluations.
2. The prevalence of complete lack of ability in such evaluations.
3. The alpha rate (Type I error rate) used by the examiner to reject or fail to reject null hypotheses.
4. The rate at which malingerers score below chance.

For example, Frederick and Denney (1998) computed the predictive power of below-chance responding on an amnesia FCT within a criminal forensic evaluation setting given prevalence of malingered amnesia = 5%, prevalence of true amnesia = 1%, overall $\alpha = 5\%$, $\alpha$ for rejection = 2.5%, and rate of below-chance responding for malingerers = 25%. Using these values and an observed *x* generating a *z*-value = –1.96, they computed a positive predictive power = 97.7%.

Bieliauskas, Fastenau, Lacy, and Roper (1997), Millis and Volinsky (2001), Mossman and Hart (1996), and Mossman (2000a, 2000b, 2003) discussed other statistical methodologies for determining the likelihood that test scores resulted from malingering. Mossman (2003) conceded that such statistical procedures were likely inaccessible to the average clinician (see also Guilmette & Giuliano, 1991). One might then wonder how accessible the discussion of such analyses would be to the average decision maker who must hear evidence about performance on malingering tests. Millis and Volinsky rightly emphasize that single test scores should never be presented to decision makers in isolation. Rogers (1997) has long promoted an approach of synthesizing clinical data to address the strength and consistency of results on psychological tests and the absence of alternative explanations for improbable events.

Perhaps the most important thing we can do with respect to FCT outcomes is to make distinctions about whether we are evaluating test scores with respect to

some empirically derived cutoff score or whether we are overtly comparing test scores to those expected by guessing. When we state that the observed score is improbable, we should be able to articulate on what basis it is improbable. If we are evaluating the probability of a test score with respect to chance, we should be clear about our hypotheses, our alpha, our test statistic, and its rejection ranges. We should be able to articulate the ways in which the observed score could have been generated and we should be able to assign some general confidence to each of those possibilities. Finally, if we observe an improbable event, we should be able to place it in context, discussing why it impresses us or why it fails to impress us given all the information available to us. Future research should investigate those ways in which decision makers most benefit from information about improbable events.

## APPENDIX

There are a couple of standard ways to compute the probability of a score on a forced-choice testing (FCT) procedure. The first is to compute the probability by standard binomial expansion:

$$\text{probability of } k \text{ correct answers out of } N \text{ trials} = \frac{N!}{k!(N-k)!} p^k q^{(N-k)},$$

where $p$ is the probability of a correct response and $q$ is the probability of an incorrect response. The probability we generally derive by this procedure, the probability of $k$ responses or fewer, is actually the sum of all the probabilities from 0 to $k$. Consequently, the probability of observing $Np$ or less is not .5. For example, the probability of observing 49 correct choices or fewer out of 100 FCT trials is .46. The probability of observing exactly 50 correct choices is .08. So the probability of observing 50 or fewer correct responses is .46 + .08 = .54.

When $p = q$, which occurs when there are only two answer choices, this formula becomes:

$$\text{probability of } k \text{ correct answers out of } N \text{ trials} = \frac{N!}{k!(N-k)!} p^N.$$

Another way to compute the probability is to compute the $z$-score for the number of correct responses (by convention, we will now refer to the number of correct responses as $x$):

$$z = \frac{\left[(x \pm .5) - Np\right]}{\sqrt{Npq}}.$$

The $z$ score can be converted to a probability value. The adjustment to $x$ (adding .5 when $x < Np$; subtracting .5 when $x > Np$) is made to correct for lack of continuity because the binomial distribution involves a discrete variable (Siegel, 1956).

## NOTES

1. These precise characteristics of *Np* are restricted to instances when *N* is even and *Np* is an integer. When *N* is odd, some of our statements about *Np* must be modified slightly. For example, when *N* is odd, two scores on either side of *Np* are the modes (we cannot observe half responses).

2. A third, remote, potential influence is nonwillful, or unconscious, choosing of incorrect responses (see Frederick, Carter, & Powel, 1995).

3. Stated more precisely, we are interested in the probability that $X \leq x$ when $p = .5$ *or* $X \geq x$ when $p = .5$. X is the random variable representing the number of correct responses associated with the performance before the trials are administered.

4. Again, by idiosyncrasy, we *do not* mean anything about the content of the item that generates differential rates of response for cooperative test-takers based on knowledge of culture or ease of solubility, such as in "rarely missed items" (see Killgore & DellaPietra, 2000). Therefore, an idiosyncratic characteristic of a test item that would prove problematic to the hypothesis $p = .5$ is observed in a test item that generates differential rates of A or B answer choices without respect to any contribution of test-takers.

5. Because a null hypothesis must always contain some form of equality, we note that this construction is inherently incorrect.

6. Note that this is exactly the primary strategy of popular tests such as the Portland Digit Recognition Test (PDRT) and the Test of Memory Malingering (TOMM). The distinction is that researchers have empirically established that cut scores that are above chance are effective at identifying malingering. The strategy is not employing a comparison to chance.

7. Frederick and Denney (1998) showed that variability in $p$ among trials, as long as overall $p = .5$, actually results in more conservative decision making with respect to $H_o$: $p_i = .5$.

## REFERENCES


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

Bieliauskas, L. A., Fastenau, P. S., Lacy, M. A., & Roper, B. L. (1997). Use of odds ratio to translate neuropsychological scores into real-world outcomes: From statistical significance to clinical significance. *Journal of Clinical and Experimental Neuropsychology*, *19*, 889-896.

Bigler, E. D. (1990). Neuropsychology and malingering: Comment on Faust, Hart, and Guilmette (1988). *Journal of Consulting and Clinical Psychology*, *58*, 244-247.

Binder, L. M. (1990). Malingering following mild head trauma. *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. (2002). The Portland Digit Recognition Test: A review of validation data and clinical use. *Journal of Forensic Neuropsychology*, *2*, 27-41.

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

Brady, P. B., & Lind, D. L. (1961). Experimental analysis of hysterical blindness. *Archives of General Psychiatry*, *4*, 331-339.

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

Elwood, R. W. (1993). Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review*, *13*, 409-419.

Etcoff, L. M., & Kampfer, K. M. (1996). Practical guidelines in the use of symptom validity and other psychological tests to measure

malingering and symptom exaggeration in traumatic brain injury cases. *Neuropsychological Review*, *6*, 171-201.

Faust, D., & Auckley, M. A. (1998). Did you think it was going to be easy? Some methodological suggestions for the investigation and development of malingering detection techniques. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 1-54). New York: Plenum.

Faust, D., & Guilmette, T. J. (1990). To say it's not so doesn't prove that it isn't: Research on the detection of malingering (reply to Bigler). *Journal of Consulting and Clinical Psychology*, *58*, 248-250.

Frederick, R. I. (1997). *The Validity Indicator Profile*. Minnetonka, MN: NCS Assessments.

Frederick, R. I. (2000). Mixed group validation: A method to address the limitations of criterion group validation in research on malingering detection. *Behavioral Sciences and the Law*, *18*, 693-718.

Frederick, R. I. (2003). *The Validity Indicator Profile* (2nd ed.). Minnetonka, MN: NCS Pearson.

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.

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

Frederick, R. I., Crosby, R. D., & Wynkoop, T. F. (2000). Performance curve classification of invalid responding on the Validity Indicator Profile. *Archives of Clinical Neuropsychology*, *15*, 281-300.

Frederick, R. I., & Denney, R. L. (1998). Minding your "*p*s and *q*s" when conducting forced-choice recognition tests. *Clinical Neuropsychologist*, *12*, 193-205.

Grosz, H. J., & Zimmerman, J. (1965). Experimental analysis of hysterical blindness: A follow-up report and new experimental data. *Archives of General Psychiatry*, *13*, 255-260.

Guilmette, T. J., & Giuliano, A. J. (1991). Taking the stand: Issues and strategies in forensic neuropsychology. *Clinical Neuropsychologist*, *5*, 197-219.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston.

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. (1994). Application of a forced-choice memory procedure designed to detect experimental malingering. *Archives of Clinical Neuropsychology*, *9*, 437-450.

Killgore, W. D. S., & DellaPietra, L.(2000). Using the WMS-III to detect malingering: Empirical validation of the Rarely Missed Index (RMI). *Journal of Clinical and Experimental Neuropsychology*, *22*, 761-771.

Lees-Haley, P. R., Dunn, J. T., & Betz, B. P. (1999, Fall). Test review: The Victoria Symptom Validity Test. *American Psychology-Law Society Newsletter*, *19*(3), 12-16.

Meyers, J., & Volbrecht, M. (1998). Validation of reliable digits for detection of malingering. *Assessment*, *5*, 303-307.

Millis, S. R., & Volinsky, C. T. (2001). Assessment of response bias in mild head injury: Beyond malingering tests. *Journal of Clinical and Experimental Neuropsychology*, *23*, 809-828.

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology*, *24*, 1094-1102.

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

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

Mossman, D. (2003). Daubert, cognitive malingering, and test accuracy. *Law and Human Behavior*, *27*, 229-249.

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

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., Fausti, S. A., & 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.

Reitan, R., & Wolfson, D. (1996). The question of validity of neuropsychological test scores among head-injured litigants: Development of a Dissimulation Index. *Archives of Clinical Neuropsychology*, *11*, 573-580.

Reynolds, C. R. (1998). Common sense, clinicians, and actuarialism in the detection of malingering during head injury litigation. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 261-286). New York: Plenum.

Rogers, R. (1997). Current status of clinical methods. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 373-397). New York: Guilford.

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.

Rosenfeld, B., Sands, S. A., & van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology*, *15*, 349-359.

Ruiz, M. A., Drake, E. B., Glass, A., Marcotte, D., & van Gorp, W. G. (2002). Trying to beat the system: Misuse of the Internet to assist in avoiding the detection of psychological symptom dissimulation. *Professional Psychology: Research & Practice*, *33*, 294-299.

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.

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *Clinical Neuropsychologist*, *13*, 545-561.

Slick, D. J., Tan, J. E., Strauss, E. H., Mateer, C., Harnadek, M., & Sherman, E. (2003). Victoria Symptom Validity Test scores of patients with profound memory impairment: Nonlitigant case studies. *Clinical Neuropsychologist*, *17*, 390-394.

Tan, J. E., Slick, D. J., Strauss, E., & Hultsch, D. F. (2002). How'd they do it? Malingering strategies on symptom validity tests. *Clinical Neuropsychologist*, *16*, 495-505.

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.

Thompson, G. B. (2002). The Victoria Symptom Validity Test: An enhanced test of symptom validity. *Journal of Forensic Neuropsychology*, *2*, 43-67.

Tombaugh, T. N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, *9*, 260-268.

Tombaugh, T. N. (2002). The Test of Memory Malingering (TOMM) in forensic psychology. *Journal of Forensic Neuropsychology*, *2*, 69-96.

**Richard I. Frederick**, PhD, is staff psychologist in the Department of Psychology, U.S. Medical Center for Federal Prisoners, Springfield, Missouri. He received his doctorate in clinical psychology from Oklahoma State University in 1986 and his MS in mathematics from Texas A&M University in 2004.

**F. Michael Speed**, PhD, is professor of statistics in the Department of Statistics at Texas A&M University. He received his doctorate in statistics from Texas A&M University in 1969.