# The Influence of Test Characteristics on the Detection of Aberrant Response Patterns

**Steven P. Reise**
**University of California, Riverside**

**Allan M. Due**
**University of Minnesota**

Statistical methods to assess the congruence between an item response pattern and a specified item response theory model have recently proliferated. This "person fit" research has focused on the question: To what extent can person-fit indices identify well-defined forms of aberrant item response? This study extended previous person-fit research in two ways. First, an unexplored model for generating aberrant response patterns was explicated. The data-generation model is based on the theory that aberrant item responses result in less psychometric information for the individual than predicated by the parameters of a specified response model. Second, the proposed response aberrancy generation model was implemented to investigate how the aberrancy detection power of a person-fit statistic is influenced by test properties (e.g., the spread of item difficulties). Results indicated that detecting aberrant response patterns was especially problematic for tests with less than 20 items, and for tests with limited ranges of item difficulty. An applied consequence of these results is that certain types of test designs (e.g., peaked tests) and administration procedures (e.g., adaptive tests) potentially act to limit the detection of aberrant item responses. *Index terms: aberrancy detection, IRT, person fit, response aberrancy, $Z_L$* index.

Since the introduction of item response models (Lord & Novick, 1968), researchers routinely have investigated statistical indices that assess the congruence between item responses aggregated across examinees and a specified item response theory (IRT) measurement model. This domain of empirical exploration has been termed "item fit" research, and the corresponding statistical indices are referred to as item-fit indices (e.g., McKinley & Mills, 1985; Yen, 1981). Within the last decade, statistical indices have proliferated that assess the degree to which an examinee's responses aggregated across items are congruent with an IRT model (e.g., Levine & Rubin, 1979; Tatsuoka 1984). These indices of person-to-model congruence are labeled "person-fit" indices.

Person-fit indices have diverse applications. For example, Harnisch (1983) discussed educational diagnostics, Tatsuoka and Tatsuoka (1982) showed their use in clarifying dimensionality, and van der Flier (1982) presented an application in cross-cultural research. One particularly appealing future application may lie in the identification of response patterns that are so incongruent with the measurement model that the predictive validity of an individual's test score can be seriously questioned. Detecting nonpredictive test scores accurately would be highly advantageous; to this end, it would be desirable if scores on a person-fit index could be equated with the corresponding test score's lack of predictive utility. At present, however, the use of person-fit indices for this purpose has not been empirically established.

This research had three objectives. The first was to present a person-fit index, called $Z_L$, which was developed by Drasgow, Levine, and Williams (1985). This so-called "practical" index has received substantial attention in the empirical research literature (e.g., Birenbaum, 1985; Drasgow, Levine, & McLaughlin, 1987; Gafni,

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 15, No. 3, September 1991, pp. 217-226*
*© Copyright 1991 Applied Psychological Measurement Inc.*
*0146-6216/91/030217-10$1.75*

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

1987; Molenaar & Hoijtink, 1990). The second objective was to discuss a potential model for conceptualizing response aberrancy and to propose a data simulation method based on this model. The third objective was to investigate, by means of the proposed simulation method, how the detection power of the $Z_L$ statistic is influenced by several test properties.

## The $Z_L$ Index

Numerous researchers (e.g., Levine & Rubin, 1979) have proposed statistics to detect whether an item response pattern is congruent with a specified measurement model. However, not all person-fit indices presume an IRT measurement model. For example, Harnisch and Linn (1981) examined properties of several non-IRT indices. An important finding of their research was that the distributions of person-fit index scores depended on examinee scale score. An association between scale score and a person-fit index limits the interpretability of the person-fit statistic (Molenaar & Hoijtink, 1990). Recognizing this, researchers (e.g., Tatsuoka, 1984) have developed IRT-based person-fit indices that are standardized with respect to trait level. These standardized indices eliminate the confounding influence of examinee trait level on the interpretation of the person-fit statistic.

In this study, the power of a standardized person-fit index ($Z_L$; Drasgow et al., 1985) was examined. To understand how this statistic functions, consider the three-parameter IRT model in Equation 1:

$$P|\theta = c + \frac{(1-c)}{1 + \exp[-a(\theta - b)]} \quad , \qquad (1)$$

where $\theta$ is the continuous latent trait,
$a$ is the item discrimination,
$b$ is the item difficulty,
$c$ is the guessing parameter, and
$P|\theta$ is the probability of a correct item reresponse conditional on $\theta$.

Given that an examinee has completed a set of dichotomously-scored test items with specified parameters, $\theta$ can be estimated by maximizing the log-likelihood function in Equation 2:

$$L|\theta = \sum\{U[\ln(P|\theta)] + (1-U)[\ln(Q|\theta)]\} \ . \qquad (2)$$

In this equation, $U$ represents the 0 (incorrect) or 1 (correct) response, $Q|\theta = 1 - P|\theta$ is the conditional probability of an incorrect response, and the summation is performed over items. The distribution of $L|\theta$ is not independent of $\theta$ (see Drasgow et al., 1985, p. 73). If this confound were ignored, the values of $L|\theta$ could be used as a person-fit index. However, the confound between $\theta$ and $L|\theta$ can be limited by using the approximate standardizing relations given in Drasgow et al. (1985).

The expected value of $L|\theta$ is

$$E(L|\theta) = \sum\{[P|\theta]\,[\ln(P|\theta)] + [Q|\theta] \times [\ln(Q|\theta)]\} \qquad (3)$$

and the variance is given by

$$V(L|\theta) = \sum\{(P|\theta)\,(Q|\theta)[\ln(P|\theta/Q|\theta)]^2\} \ . \qquad (4)$$

Taking the summations in Equations 3 and 4 over items and putting the terms together, the standardized person-fit index is

$$Z_L = \frac{[L|\theta - E(L|\theta)]}{V(L|\theta)^{1/2}} \ . \qquad (5)$$

In practice, of course, an estimate of trait level ($\hat{\theta}$) must be inserted in all equations.

The above standardization of the $L|\theta$ allows examinees at different $\theta$ levels to be compared on the basis of their $Z_L$ aberrancy scores. $Z_L$ scores have an expected value of 0 and a variance of 1.0, conditional on $\theta$, under the null hypothesis that the response vectors were generated by the specified model. It is also assumed that the distribution of $Z_L$ scores is normal, conditional on $\theta$, in populations conforming to the model. Recent research (Molenaar & Hoijtink, 1990) questioned the appropriateness of the normality assumption for the $Z_L$ null distribution, particularly because $\theta$ must be estimated with real data. Yet the success of the normal null distribution in other research (Drasgow et al., 1987; Drasgow et al.,

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

1985; Reise, 1990) supports its use as practical and adequate for the present purposes.

Negative $Z_L$ scores indicate relatively unlikely response patterns (i.e., inconsistent respondents), and positive $Z_L$ scores indicate patterns that are more likely than the probabilistic IRT model predicts (i.e., hyperconsistent respondents). The statistical significance of an observed $Z_L$ value is computed using the standard normal distribution. The null hypothesis is that the response pattern was generated from a specified IRT model.

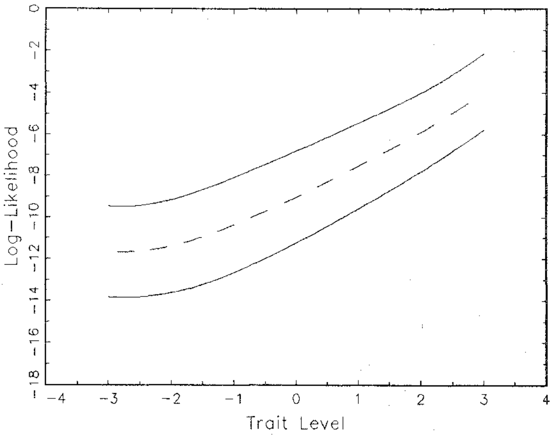
Figure 1 provides a graph of the expected values of $L|\theta$ and the corresponding ±1 standard deviation bands, using Equations 3 and 4 for an example test. The 21-item test consisted of items with $a = 1.5$, $b = U(-3,3)$, and $c = .2$. Interpretation of Figure 1 indicates that if an examinee's item responses are generated from the specified IRT model, there is a .68 probability (the area between ±1 $Z$ in the standard normal distribution) that the observed $L|\theta$ will lie between the upper and lower lines. Thus $Z_L$ values are interpreted like $Z$ scores, and the likelihood of the observed response pattern being generated from the specified IRT model is tested with the standard normal distribution.

## Models of Response Aberrancy

Two types of response patterns are technically incongruent with the probabilistic IRT model: inconsistent and hyperconsistent patterns. Although both are interesting, the concern here and in most previous fit research was with identifying inconsistent response patterns (i.e., the types of patterns that have a lower-than-expected $L|\theta$ and result in a negative $Z_L$). Consequently, the term "aberrancy" refers here to responses that result only in negative $Z_L$ values.

There are many potential causes of response aberrancy. Nevertheless, if a researcher seeks to identify individuals whose item response patterns are not tenable given a specified IRT model, aberrant item response behaviors must be defined and the manner in which their manifestation affects a quantifiable property of the test protocol must be proposed. A distinction is made here between two approaches—statistical and information-based—to defining and quantifying aberrancy. Although these approaches share properties, maintaining their distinction has consequences for the way in which response aberrancy is understood in practice and is simulated in research.

**Figure 1**
The Expected Value of L|θ and ±1 SD as a Function of θ for the Example Test



## The Statistical Approach

At the core of the statistical approach is the notion that a response is aberrant to the extent that it is of low probability. The lower the probability of the response determined by the IRT model parameters, the more aberrant the response. This statistical perspective is clearly evident in the $Z_L$ index. $Z_L$ quantifies the relative decrease in the value of $L|\theta$ caused by a response that is unlikely, given the model. Clearly from Equations 3 and 4, the lower the probability of the observed response, the more that response will influence the $Z_L$ statistic. A correct guess to a difficult item, for example, will have a greater impact on $Z_L$ when the examinee's $\theta = -2.0$ than when the examinee's $\theta = 2.0$.

When researchers (e.g., Birenbaum, 1986; Drasgow et al., 1987; Gafni, 1987) have investigated the aberrancy detection power of $Z_L$ (i.e., the validity of $Z_L$), they typically have begun with a set of model-conforming response patterns. The model-fitting patterns were then manipulated to

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

mimic clearly defined and hypothetically identifiable "forms" of aberrancy. One common procedure, for example, has been to change various proportions of a low θ examinee's item responses from incorrect to correct. This manipulation mimics the appearance of the response patterns of cheaters. The $Z_L$ index is then computed and the detectability of simulated cheating behavior is explored. Equivalently, the validity of $Z_L$ for detecting cheating is investigated. This simulation approach is rational, yields useful results, and was not a subject of dispute in this research.

## The Information-Based Approach

The information-based perspective, however, offers an alternative model of aberrancy that might have potential value in research and applied situations. Central to the information-based perspective is the notion that an aberrant response is one that provides less psychometric information (Lord, 1980) for estimating θ than would be predicated by the parameters of a specified IRT model. This perspective is reasonable because manifestation of defined aberrancy-causing factors (e.g., cheating) should result in item responses for which the item information that results from the specified IRT model is positively biased. Given this definition of aberrancy, it follows that the information-based approach is concerned with quantifying the decrease in test information from that expected given a specified IRT model that is due to aberrant responses.

Because test information depends on the values of the $a$ parameter (Lord, 1980), the information-based approach requires a model of item responding in which items are differentially discriminating for different individuals. One such model, which is similar to the generalized response aberrancy model in Strandmark and Linn (1987), is

$$P|\theta = c + \frac{(1 - c)}{1 + \exp[-(a_P a_I)(\theta - b)]} \quad , \qquad (6)$$

where $a_I$ is the item discrimination, and $a_P$ is the aberrancy level parameter.

Equation 6 differs from the IRT model in Equation 1 in that (1) the $a$ parameter is indexed by $a_I$ in order to clearly indicate that it is the slope parameter from a three-parameter item response function; and (2) the $a_P$ person parameter multiplicatively weights the $a_I$ item parameter. Note that the $a_P$ parameter is a constant across items for a given examinee.

The $a_P$ parameter implies that the test information for the model in Equation 6, summed over items, becomes

$$I|\theta = \Sigma\left[\frac{(a_P a_I)^2 (Q|\theta)(P|\theta - c)^2}{(1 - c)^2}\right] \quad . \qquad (7)$$

Given a set of $a_I$, $b$, and $c$ parameters from a specified three-parameter model, the information provided by the model can be manipulated in a simulation by changing the values of $a_P$ for different individuals. Thus, systematic changes in $a_P$ can be made to generate response patterns that are congruent with the hypothesis that the test provides less information than predicated by the specified model. When $a_P = 1.0$, the model reduces to Equation 1. When $a_P = 0.0$, the item provides no information with respect to estimating θ.

To demonstrate how the manipulation of $a_P$ affects values of the $Z_L$ index, consider the test parameters described for Figure 1. Using these parameters in Equation 6, 1,000 response vectors were simulated for examinees with $\theta = 0.0$ at each of 11 levels of $a_P$. Specifically, $a_P$ was varied from 1.0 to 0.0 in .10 intervals. For each response pattern, $Z_L$ was computed based only on the three-parameter model. Hence, the response vectors were simulated with Equation 6, but were then tested for fit with the model in Equation 1. Results are shown in Table 1.

It is evident from these data that as $a_P$ decreases from 1.0, marked decreases occur in mean $Z_L$; the $Z_L$ index apparently is sensitive to manipulations of the $a_P$ parameter. This results from the fact that the expected value of the second derivative of the log-likelihood function is manipulated directly by the $a_P$ parameter in Equation 6 . The relative height of the likelihood

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

**Table 1**
$Z_L$ Distributions for 1,000 Examinees with $\theta = 0.0$ at Each of 10 $a_P$ Levels

| | $Z_L$ | |
|---|---|---|
| $a_P$ | Mean | SD |
| 1.0 | .01 | .99 |
| .9 | -.22 | 1.13 |
| .8 | -.33 | 1.82 |
| .7 | -.71 | 1.28 |
| .6 | -.99 | 1.40 |
| .5 | -1.50 | 1.59 |
| .4 | -2.24 | 1.85 |
| .3 | -2.98 | 1.92 |
| .2 | -3.82 | 2.18 |
| .1 | -4.99 | 2.26 |
| 0.0 | -6.31 | 2.46 |

functions (i.e., the property that $Z_L$ was designed to quantify) is affected only indirectly.

The drawback to simulating aberrancy under Equation 6 is that no specific aberrant behaviors (e.g., cheating) are modeled. There are two benefits of using Equation 6, however. First, if $a_P$ is viewed as a person parameter (Lumsden, 1977; Strandmark & Linn, 1987), the model supplies an interesting framework for conceptualizing aberrancy. That is, aberrancy can be viewed as resulting from the relative failure of a test item to discriminate the θ level for the examinee, which results in less informative measurement for that examinee. Second, Equation 6 is potentially useful in research for investigating an important model characteristic—specifically, how well the $Z_L$ index (or other IRT indices of person fit) identifies responses that are less informative than the IRT model predicates.

### The Power of $Z_L$ to Detect Changes in $a_P$

Previous investigations of $Z_L$ have studied how well it can detect well-defined forms of aberrant item response behavior. The research has typically been set within the context of a single lengthy (e.g., 85-item) test. Of concern here was the power of $Z_L$ to detect response aberrancy as a function of three test characteristics: (1) the number of items, (2) the spread of the $b$ parameter, and (3) the value of the $c$ parameter. To create aberrant response patterns, the $a_P$ parameter was manipulated in simulations using Equation 6. However, the $Z_L$ values for all response patterns were computed based only on the three-parameter IRT model. Hence, the null hypothesis was that the response pattern fit the three-parameter model, but the data were simulated under the alternative hypothesis that the response vectors were generated from a less informative model.

## Method

### General Simulation Algorithm

Three sets of monte carlo simulations were performed to address the $Z_L$ power issue. One test characteristic in each set was systematically manipulated, and the detectabilities of several levels of aberrancy were determined. Each simulation followed a similar procedure. First, $a_I$, $b$, and $c$ parameters were specified for a three-parameter model. In addition, three values for $a_P$ were specified: 1.0, .5, and 0.0. These $a_P$ values are referred to as the three "aberrancy levels." Based on the specified parameters, 1,000 response vectors were generated at each of 11 points on the θ continuum using the model in Equation 6. The 11 true θ values ranged from -2.5 to 2.5 θ in .5 intervals. Thus, 33,000 response vectors were simulated (11 θ levels × 3 $a_P$ levels × 1,000 simulees) for each specified set of item parameters.

$Z_L$ was computed for each vector based on a model without the $a_P$ parameter (i.e., $Z_L$ was computed using only the parameters of the standard three-parameter model). Each $Z_L$ value was then tested for significance using the normal distribution as the null. The critical value was specified as $Z_L = -1.65$, which represents the one-tailed $\alpha = .05$ error rate. $Z_L$ values significant at the $\alpha = .05$ level were considered to be "hits." The detection hit rates within each condition were found by counting the number of hits and then dividing by 1,000.

The first level of $a_P$ (1.0) produced response patterns that were in accordance with Equation 1; thus, these conditions represented the null con-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

dition (i.e., when data are generated and tested for fit under the same three-parameter model). The second manipulation ($a_P = .5$) resulted in response vectors from a model in which the $a_I$ parameters were reduced in half, which reduced test information. Finally, the third condition ($a_P = 0.0$) produced response vectors that provided no psychometric information.

As mentioned above, $Z_L$ was computed based on the originally specified three-parameter model (i.e., using only $a_I$, $b$, and $c$) for each simulated response vector. No estimate of θ was obtained; instead, the true generating θ value was used in all computations of $Z_L$. Of course, the use of the true θ produced higher detection hit rates than if an estimate had been used. The central concerns in this study, however, were the effects of test characteristics on the power of $Z_L$. Estimating θ might have detracted from the interpretability of the simulations, in addition to adding potentially another facet to the design.

### Test Characteristics

The effects of test length on detection power were studied by specifying item parameters for a 7-item test. The specified parameters were: $a_I = 1.5$, $b$ $U(-3,3)$ at unit intervals, and $c = .2$. The same parameter distributions were then used to create three additional tests with 21, 35, and 49 items.

Item difficulty range effects were studied by specifying $b$ parameters for five tests. The first test was created by taking the item parameters used in the 49-item conditions specified above. Then a second test was specified by collapsing the $b$ parameters by factors of .5 toward 0.0; of course, $b = 0.0$ were unchanged. Thus, $b$ parameters ranged from −3.0 to 3.0 in the first test and from −2.5 to 2.5 in the second test. The third test was created by moving the $b$ parameters in the second test toward 0 by a factor of .5. This process of creating a new test from the previous test was continued until there were five tests. The fifth test had $b$ parameters ranging from −1.0 to 1.0.

Because the $a_I$ were never changed between tests, the manipulations resulted in each test having the same total amount of information across the θ range, but that information was concentrated toward the middle of the θ scale ($\theta = 0.0$) as each new test was created. Test information functions were computed for the five tests in order to judge this effect. Power analyses were then conducted for each of the five tests.

To isolate the effects of the guessing parameter, five 49-item tests were specified. For each test, $a_I = 1.5$ and the $b$ were $U(-3,3)$. The $c$ parameters were then specified to equal 0.0, .05, .10, .15, and .20, respectively, across the five tests.

## Results

### Test Length

The hit rate results for the four test-length conditions are shown in Table 2. Hit rates were higher than .05 under several conditions when $a_P = 1.0$ (i.e., when data were generated according to the three-parameter model). For a 7-item test and $\theta = 2.5$, for example, 10% of simulees were identified as aberrant. These over-rejection results, which appear to have occurred more frequently at the lower test lengths and the higher θ levels, indicate that $Z_L$ rejected the null hypothesis of fit at a rate higher than would be expected under $\alpha = .05$. Moreover, there were only two cases in the $a_P = 1.0$ conditions in which the hit rates were below .05. This bias of $Z_L$ toward rejection was not substantial, however, and does not seriously affect the interpretation of the remaining results in Table 2.

For the $a_P = .50$ and $a_P = 0.0$ conditions, higher detection hit rates were associated with greater aberrancy levels. More importantly, hit rates increased rather substantially as test length increased. Within the $a_P = .50$ conditions, for example, hit rates approximately doubled between the 7- and 49-item simulations. In the $a_P = 0.0$ conditions, detection was almost perfect for tests longer than 21 items. Yet the hit rates reached a low of approximately 58% in the low θ levels at seven items. Apparently, at least seven items are needed in order to have a greater than 50/50

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

**Table 2**
$Z_L$ Hit Rates for Increasing Test Lengths and Three Levels of Aberrancy ($a_P$) at Selected Values of θ

| | Test Length | | | |
|---|---|---|---|---|
| θ | 7 | 21 | 35 | 49 |
| $a_P$ = 1.0 | | | | |
| -2.50 | .05 | .05 | .05 | .05 |
| -2.00 | .06 | .06 | .06 | .05 |
| -1.50 | .08 | .05 | .06 | .06 |
| -1.00 | .05 | .06 | .05 | .05 |
| -.50 | .07 | .06 | .05 | .06 |
| 0.00 | .06 | .05 | .06 | .04 |
| .50 | .09 | .06 | .06 | .05 |
| 1.00 | .06 | .06 | .06 | .07 |
| 1.50 | .05 | .04 | .06 | .07 |
| 2.00 | .06 | .07 | .06 | .06 |
| 2.50 | .10 | .07 | .07 | .07 |
| $a_P$ = .5 | | | | |
| -2.50 | .10 | .15 | .20 | .25 |
| -2.00 | .14 | .18 | .25 | .31 |
| -1.50 | .18 | .26 | .35 | .41 |
| -1.00 | .20 | .33 | .42 | .50 |
| -.50 | .22 | .39 | .49 | .59 |
| 0.00 | .26 | .42 | .56 | .63 |
| .50 | .28 | .44 | .58 | .66 |
| 1.00 | .28 | .45 | .61 | .68 |
| 1.50 | .27 | .50 | .61 | .73 |
| 2.00 | .30 | .47 | .60 | .71 |
| 2.50 | .33 | .46 | .63 | .71 |
| $a_P$ = 0.0 | | | | |
| -2.50 | .58 | .94 | .99 | 1.00 |
| -2.00 | .57 | .92 | .99 | 1.00 |
| -1.50 | .58 | .94 | .99 | .99 |
| -1.00 | .64 | .95 | .99 | 1.00 |
| -.50 | .70 | .95 | .99 | 1.00 |
| 0.00 | .75 | .96 | .99 | 1.00 |
| .50 | .79 | .98 | .99 | 1.00 |
| 1.00 | .80 | .99 | 1.00 | 1.00 |
| 1.50 | .84 | .99 | 1.00 | 1.00 |
| 2.00 | .89 | .99 | .99 | 1.00 |
| 2.50 | .91 | .99 | 1.00 | 1.00 |

chance of identifying aberrancy in its most extreme form (i.e., when $a_P$ = 0.0 is suspected) at low θ levels.

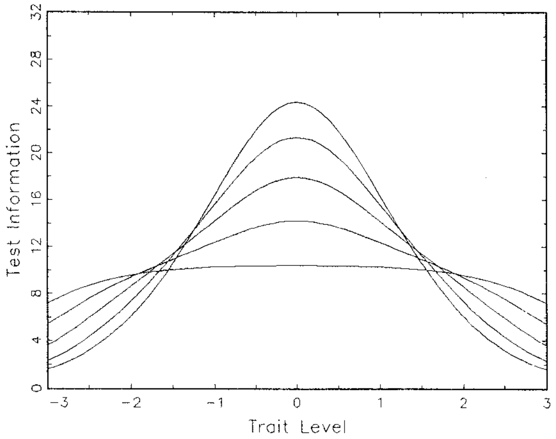
An obvious feature of Table 2 is that hit rates were larger at the positive end of the θ continuum than at the negative end. As mentioned above, the resultant $Z_L$ values were more extreme with lower probabilities of response. This could occur for individuals at low θ levels because the guessing parameter makes it difficult to respond unexpectedly; there always exists at least a .2 probability of item endorsement (the less-likely event) for these examinees. The potential contribution of an aberrant response to the $Z_L$ index is not symmetric in the three-parameter model, which has a much greater contribution at the high end of the θ continuum than at the low end. Consequently, the detection is greater for positive values of θ.

## Spread of *b* Parameters

Test information functions were computed for the five tests in order to evaluate the effects of the spread of the *b* parameters; these curves are displayed in Figure 2. Power analyses were then conducted for each of the five exams (see Table 3).

The results for the $a_P$ = 1.0 conditions show that $Z_L$ rejected fit at approximately the .05 level across the various simulations. This indicates adherence to the predicted rejection rates under the null distribution. When $a_P$ = .5, two results were clear. First, hit rates within tests were affected strongly by θ level. When information was 10.4, for example, the hit rate was .39 for θ = -1.5, and it was .70 when θ = 1.5. Second, hit rates increased for θ values greater than 1.5 or less than -1.5 as the tests became peaked in information.

**Figure 2**
Test Information Functions for Five Tests That Differ Only in Item Difficulty Distribution



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

**Table 3**
$Z_L$ Hit Rates for Increasingly Peaked Tests and Three Levels of Aberrancy ($a_P$) at Selected Levels of θ

| | Information at θ = 0.0 | | | | |
|---|---|---|---|---|---|
| θ | 10.4 | 14.2 | 17.9 | 21.3 | 24.3 |
| $a_P$ = 1.0 | | | | | |
| -2.5 | .05 | .05 | .05 | .05 | .05 |
| -2.0 | .05 | .04 | .05 | .05 | .04 |
| -1.5 | .03 | .05 | .04 | .05 | .06 |
| -1.0 | .06 | .06 | .05 | .05 | .04 |
| -.5 | .05 | .05 | .04 | .06 | .06 |
| 0.0 | .05 | .04 | .05 | .04 | .05 |
| .5 | .05 | .05 | .05 | .05 | .06 |
| 1.0 | .06 | .06 | .06 | .06 | .05 |
| 1.5 | .06 | .06 | .07 | .07 | .05 |
| 2.0 | .08 | .06 | .05 | .06 | .04 |
| 2.5 | .05 | .07 | .06 | .07 | .07 |
| $a_P$ = .5 | | | | | |
| -2.5 | .23 | .25 | .32 | .38 | .37 |
| -2.0 | .30 | .28 | .34 | .43 | .52 |
| -1.5 | .39 | .37 | .39 | .44 | .49 |
| -1.0 | .52 | .44 | .40 | .41 | .40 |
| -.5 | .58 | .52 | .43 | .37 | .24 |
| 0.0 | .64 | .57 | .47 | .38 | .22 |
| .5 | .68 | .64 | .58 | .52 | .37 |
| 1.0 | .71 | .72 | .68 | .67 | .65 |
| 1.5 | .70 | .71 | .76 | .78 | .78 |
| 2.0 | .70 | .74 | .81 | .83 | .84 |
| 2.5 | .70 | .77 | .82 | .89 | .90 |
| $a_P$ = 0.0 | | | | | |
| -2.5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| -2.0 | .99 | 1.00 | .99 | 1.00 | 1.00 |
| -1.5 | 1.00 | .99 | .99 | .99 | .99 |
| -1.0 | .99 | .99 | .98 | .98 | .97 |
| -.5 | 1.00 | .99 | .97 | .91 | .72 |
| 0.0 | 1.00 | .99 | .98 | .88 | .58 |
| .5 | 1.00 | 1.00 | .99 | .97 | .90 |
| 1.0 | 1.00 | 1.00 | .99 | 1.00 | .99 |
| 1.5 | 1.00 | 1.00 | .99 | 1.00 | 1.00 |
| 2.0 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2.5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

In contrast, hit rates decreased as the peakedness increased for θ values between -1.5 and 1.5.

The $a_P$ = 0.0 conditions were relatively uninformative due to high hit rates across the test conditions. However, note that in the extreme case of a very peaked test (information = 24.3) only 58% of aberrant patterns were identified for examinees at θ = 0.0. These findings illustrate that it is not the amount of information that allows for high detection hit rates, but rather the spread of the *b* parameters relative to the θ level that determines detection hit rates. As the item difficulties cluster around the examinee's θ level, all response probabilities approach .50. Hence, random response patterns are difficult to distinguish from normal response patterns.

### The *c* Parameter

The resultant hit rates for the *c* parameter conditions are displayed in Table 4. As in the previous analyses, the hit rates were approximately .05 within the $a_P$ = 1.0 conditions. This again illustrates a fair adherence to the null distribution rejection rates under the present conditions. Also, as for the previous analyses, in the $a_P$ = .5 and $a_P$ = 0.0 conditions, the detection power was maximized at the positive end of the θ continuum.

In the $a_P$ = .5 conditions, as *c* increased across tests, the hit rates for $Z_L$ decreased at the negative end of the θ continuum. But when the *c* parameter approached 0.0, aberrancy hit rates were detected symmetrically at both ends of the θ continuum. In addition, response aberrancy was detected with greater precision throughout the θ continuum as *c* decreased. But at the low θ extremes, hit rates were most strongly affected by levels of *c*. The $a_P$ = 0.0 conditions were simply uninformative because aberrancy was always detectable under these simulations.

Thus, just as an elevated *c* parameter decreases the amount of psychometric information available in a test, it also adversely affects the power to detect response aberrancy. Given these patterns of results, $Z_L$ is likely a more powerful detector of response aberrancy for data that fit a two-parameter model.

## Discussion

The results demonstrated that test length, spread of *b* parameters, and level of *c* all influence the ability of $Z_L$ to detect aberrant responding as operationalized in this study. Although the results indicated that $Z_L$ adhered to the expected hit rates under null conditions, they also indicated that aberrancy detection may be difficult in situations

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

**Table 4**
$Z_L$ Hit Rates for Increasing Values of *c* and Three Levels of Aberrancy ($a_P$) at Selected Values of θ

| | *c* Parameter Value | | | | |
|---|---|---|---|---|---|
| θ | 0.00 | .05 | .10 | .15 | .20 |
| $a_P$ = 1.0 | | | | | |
| −2.5 | .06 | .05 | .06 | .06 | .05 |
| −2.0 | .04 | .05 | .06 | .06 | .05 |
| −1.5 | .05 | .05 | .05 | .06 | .05 |
| −1.0 | .04 | .05 | .05 | .05 | .04 |
| −.5 | .06 | .06 | .04 | .06 | .04 |
| 0.0 | .06 | .06 | .05 | .06 | .06 |
| .5 | .05 | .07 | .04 | .06 | .06 |
| 1.0 | .05 | .05 | .06 | .06 | .06 |
| 1.5 | .04 | .06 | .05 | .05 | .06 |
| 2.0 | .03 | .06 | .06 | .05 | .08 |
| 2.5 | .05 | .07 | .06 | .06 | .07 |
| $a_P$ = .5 | | | | | |
| −2.5 | .77 | .47 | .34 | .27 | .24 |
| −2.0 | .77 | .53 | .40 | .36 | .29 |
| −1.5 | .80 | .59 | .49 | .44 | .39 |
| −1.0 | .82 | .69 | .62 | .53 | .51 |
| −.5 | .86 | .74 | .67 | .63 | .58 |
| 0.0 | .87 | .77 | .69 | .66 | .62 |
| .5 | .86 | .80 | .73 | .70 | .66 |
| 1.0 | .80 | .81 | .75 | .72 | .71 |
| 1.5 | .79 | .80 | .78 | .74 | .72 |
| 2.0 | .77 | .78 | .77 | .74 | .70 |
| 2.5 | .77 | .77 | .76 | .74 | .71 |
| $a_P$ = 0.0 | | | | | |
| −2.5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| −2.0 | 1.00 | 1.00 | 1.00 | .99 | .99 |
| −1.5 | 1.00 | 1.00 | 1.00 | .99 | .99 |
| −1.0 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| −.5 | 1.00 | 1.00 | 1.00 | 1.00 | .99 |
| 0.0 | 1.00 | 1.00 | 1.00 | 1.00 | .99 |
| .5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1.0 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1.5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2.0 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2.5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

in which it is most desirable. For example, when using the three-parameter model, aberrancy detection was limited at the low end of the θ continuum. Unfortunately, aberrant response patterns are more likely as θ values decrease, precisely the conditions under which the $Z_L$ index attained the lowest power in detecting aberrant patterns of response.

Adaptive testing is an application in which it might be difficult to identify aberrant respondents. $Z_L$ depends on a distribution of difficulty to achieve its goals. Detection of aberrant patterns, of course, increases when examinees respond aberrantly to items away from their θ level. However, note that this is contrary to the item selection goals of adaptive testing. The θ range in which individuals are measured with greatest accuracy (i.e., in which information is maximized) is also the θ range in which response aberrancy is most difficult to detect. A further ramification of this result is that classically designed peaked tests, which maximize score reliability by minimizing the range of item difficulty, will have restricted power to detect response aberrancy for examinees with θ near the average *b* parameter.

The present results lack generalizability for three reasons: (1) $Z_L$ was always computed based on θ, (2) $a_P$ was held constant across items, and (3) the item parameters (e.g., *c*) were in many cases held constant within tests. These characteristics are not expected to hold in actual testing applications. Nevertheless, the results do make explicit the effects of certain ‘‘changeable’’ properties of a test. The present results imply that once a set of test parameters is estimated under a three-parameter IRT model, monte carlo studies should be implemented to determine the power of several aberrancy detection statistics to detect aberrant response patterns under various levels of the $a_P$ parameter. Furthermore, having a range of item difficulties is of critical concern for discovering response aberrancy.

## References

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement, 45,* 523–534.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10,* 167–174.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59–79.

Drasgow, F., Levine, M. V., & Williams, E.A. (1985).

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Gafni, N. (1987). *Detection of systematic and unsystematic aberrancy as a function of ability estimate by several person-fit indices.* Unpublished doctoral thesis, University of Minnesota, Minneapolis.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20,* 191-206.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133-146.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-289.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1,* 477-482.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49-57.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75-106.

Reise, S. P. (1990). A comparison of person and item fit methods of assessing fit in IRT. *Applied Psychological Measurement, 14,* 127-137.

Strandmark, N. L., & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement, 11,* 355-370.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns. *Journal of Educational Statistics, 7,* 215-231.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13,* 267-298.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

## Acknowledgments

*The authors thank Keith Widaman for his helpful comments on an earlier draft. Also, thanks to the Editor and to two anonymous reviewers who provided suggestions and comments.*

## Author's Address

Send requests for reprints or further information to Steven P. Reise, Department of Psychology, University of California, Riverside CA 92521, U.S.A.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/