National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

A MONTE CARLO STUDY OF THE EFFECTS OF FOUR FACTORS ON THE EFFECTIVENESS OF THE LZ AND ECIZ4 APPROPRIATENESS INDICES

Gary Cole

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfillment of the requirements of the Ph.D.
degree in Education

University of Ottawa
1993

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

ISBN 0-315-93562-6

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ABSTRACT

While a test score may be valid for a group there may sometimes be reason to suspect its validity for an individual. Unusual examinee response patterns may indicate that the test may be invalid for the individual and quantitative measures called appropriateness indices have been developed to detect these unusual patterns. For a number of reasons, Lz and ECIZ4 have so far proven to be two of the most useful of these indices.

There were three purposes for this study. The first purpose was to investigate the effects of four variables on the cutoff values of the indices: the range of the distribution of the b parameter (Diff), the level of the a parameter (Disc), IRT model (Model), and sample size used to estimate item parameters (Sampsiz). The second purpose was to investigate the effects of these same variables on the detection rates for response vectors that were made spuriously high (i.e. high aberrance) and for response vectors that were made spuriously low (i.e. low aberrance). The third purpose was to determine the extent to which detection rates obtained by using cutoff values from the standard normal distribution were similar to those obtained by using cutoff values obtained by simulating non-aberrant response vectors.

Two levels were set for each of the four variables. For Diff, a broad and a narrow range of the b parameter was used. For Disc, a high and low level for the a parameter of the test items was used. For Model, the 2PL and 3PL models were used. For Sampsiz, a sample size of 1000 and 2500 was used to estimate item parameters. For each of the 16 combinations of these variables, non-aberrant as well as aberrant response vectors were simulated for a 60 item test. For the aberrant response vectors, both high and low aberrance was created by modifying ten of the test items. Detection rates were obtained at the .01, .05, and .10 false positive rates using cutoff values based on the distribution of the non-aberrant response vectors and using cutoff values based on a standard normal distribution. The simulation of each combination of conditions was replicated 90 times.

The following results were obtained:

1 The mean cutoff values for both indices were somewhat unstable over conditions and replications. However, they were more stable for ECIZ4.

2 In general, the cutoff values for both indices were significantly different from those of a standard normal distribution, but it is noted that the t - tests used to conduct the analyses were very powerful.

3 Each of the four independent variables had an effect on the cutoff values for the indices; these effects were complex and difficult to characterize.

4 The overall detection rates for both indices were quite high. For both indices, they were better for low aberrance than for high aberrance. The detection rates for ECIZ4 were generally better than the detection rates for Lz.

5 The detection rates for both indices were affected by Diff and Model. High Diff resulted in better detection rates than low Diff and the magnitude of the difference was quite large. The 2PLM produced significantly better detection rates than the 3PLM. However, while the magnitude of the difference was quite large for high aberrance it was not large for low aberrance.

6 Sampsiz had no effect on detection rates for either index.

7 For Lz high Disc produced better detection rates than low Disc, but there were no significant effects for ECIZ4.

8 The combination of the 2PLM and high Diff interacted ordinally to produce enhanced detection rates for high aberrance for both indices.

9 The detection rates obtained by using a standard normal distribution for setting cutoff values were very similar to those obtained with cutoff values based on the simulation of a non-aberrant sample. The difference between the detection rates obtained by using the two methods for setting cutoff values was less for ECIZ4.

Based on the results of the study as well as the evidence accrued from other studies four conclusions with fairly important practical consequences

were made. The first conclusion reached from this research is that it is advisable, for research purposes, to replicate conditions to avoid chance results. Detection rates varied considerably over replications. The second conclusion is that if a choice of one index must be made for practical applications, ECIZ4 should be used. The cutoff values for ECIZ4 were more stable over replications, they were closer to those of a standard normal distribution, and ECIZ4 generally had higher detection rates than Lz. The third conclusion is that there is not a great difference in detection rates for either index when cutoff scores from a standard normal distribution are used versus those set by simulation. The fourth conclusion is that the greater the range of the b parameters for the test items the higher the detection rate.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the guidance and direction which Dr. Marvin Boss provided in the preparation of this thesis. His very generous sharing of his time and expertise was invaluable. Special appreciation is reserved for the author's wife and family for their patience and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

One of the basic characteristics of a test is its validity. While numerous techniques are used to assess validity, these techniques generally involve the validity of the test for a group of individuals. However, while a test may be valid for a group there may be reason to suspect its validity for a particular individual.

The identification of individuals for whom a test has questionable validity has long been an important concern in testing. Cronbach (1946) identified some of the problems that result when a test is not valid for individuals. The test's validity as measured by the correlation with other criteria could be lowered. These invalid results would allow persons with equal knowledge, identical attitudes, or equal amounts of a personality trait to receive different scores on a test and would always reduce what Cronbach termed the test's "logical" validity. In sum, these results would interfere with the interpretation of test data.

In practical terms, one can imagine numerous situations in which serious mistakes might be avoided if individuals with invalid test results were identified. For example, if the test were being used for selection to a training program, an invalid test score could result in a great deal of frustration on the part of any trainee who is selected or rejected inappropriately. This inappropriate decision would also result in administrative problems and unnecessary cost on the part of the training institution.

It is interesting to read the research on testing from the 1940's and 1950's because it is clear that the researchers were mainly working with free response and not multiple-choice tests. Researchers tended to approach the detection of a test's validity for the individual in a different manner than at present. However, by the 1970's the multiple-choice test had become the dominant form of standardized test used in North America. The multiple-choice test offers a distinctive way of assessing whether a test is valid for

an individual because the responses of persons with a given score on a multiple-choice test are expected to follow a certain pattern. One would expect, for example, that an examinee of high ability would respond correctly to most of the easy items of a test and that an examinee of low ability would miss most of the difficult items of the test. When an individual does not follow an expected pattern of responses for a test it leads one to question the interpretation of the test results for that individual.

A number of techniques have been proposed to examine the response patterns of individuals on multiple-choice tests. The most direct way of examining an individual's response pattern is to plot it. Weiss (1973), in designing the stratified-adaptive (stradaptive) test, divided the items into difficulty levels, ordered from easiest to most difficult. A "trace line" was then plotted for the person's proportion correct number of items at each of the difficulty levels or "strata". Weiss considered that the steepness of the slope of the individual's trace line was an indicator of the "consistency" of the person's responses, while the point of inflection of the line was proposed as the indicator of the person's ability on the trait.

Lumsden (1977, 1980) also proposed plotting the individual's responses on a test by ordering the items of a test by difficulty on the x-axis, with the proportion of items answered correctly at the different difficulty levels shown along the y-axis. Lumsden called this plot a "Person Characteristic Curve" (PCC). Lumsden postulated that PCC's vary for different individuals depending on the individual's "person reliability". Flat PCC's are characteristic of persons who have large fluctuations in their ability to answer items correctly (i.e. have low person reliability).

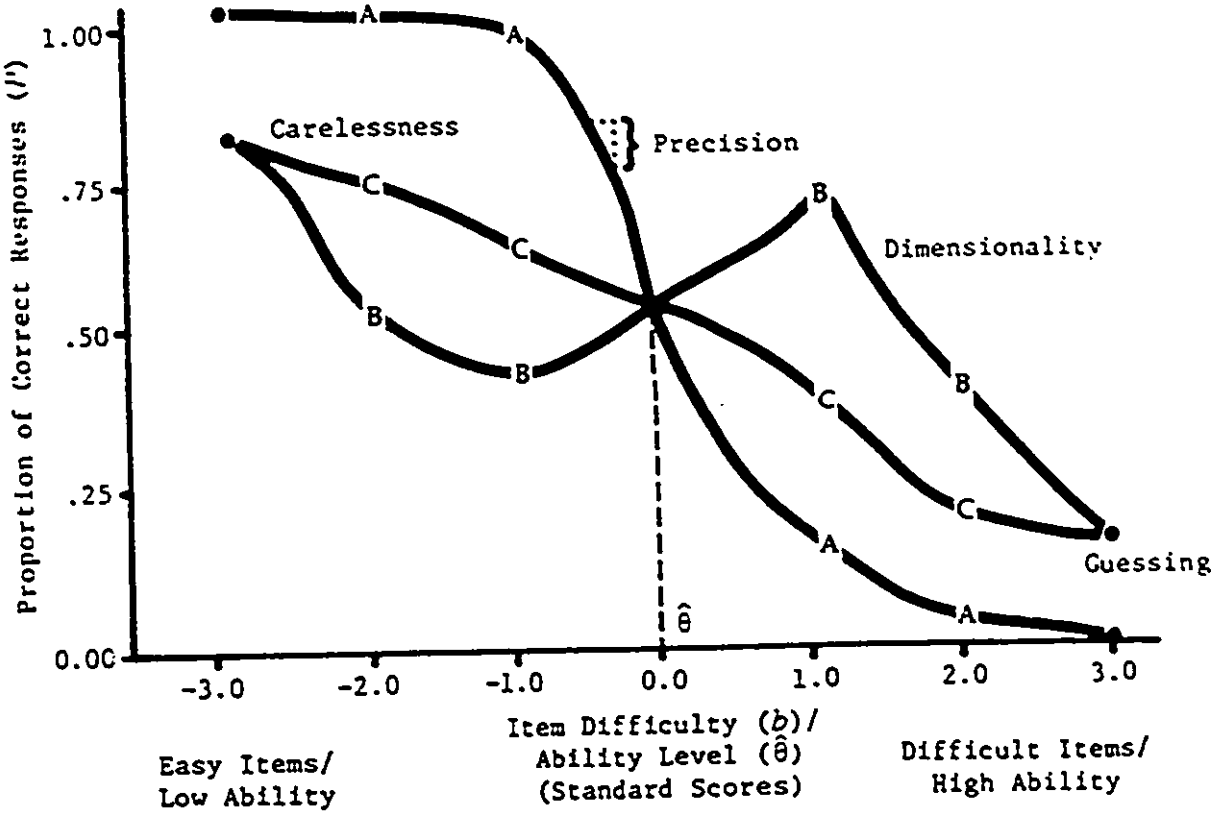
Since persons might differ in the degree to which their ability level fluctuates, Lumsden suggested that an added dimension might be necessary to fully characterize the person's ability. This person reliability characteristic would have important consequences. As Lumsden pointed out, while extreme oscillation in ability may be acceptable or even desirable in certain creative activities such as in the design of an airplane, the same

characteristics would be disastrous for flying an airplane. It is not perfectly clear whether Lumsden was associating this characteristic with a general trait or only with the ability measured by the test.

With Lumsden's model the assumption is made that PCC's all have the same form (normal ogive) and the slope of the curve is considered an adequate summary statistic for comparing person reliabilities (how much the person is likely to deviate in ability on a single test administration). Trabin and Weiss (1979) have plotted person characteristic curves based on the assumptions of item response theory (IRT). The PCC's generated by Trabin and Weiss have the same shape as Lumsden's, a normal ogive, and can be used to compare observed and expected curves to ascertain if any patterns are unusual. For unusual patterns, more detailed analyses could indicate if guessing, dimensionality, etc. were the problem.

This type of analysis is illustrated in Figure 1 which shows the PCCs for three hypothetical persons who each obtained the same score (they correctly answered 50% of the items) on a test (from Trabin and Weiss, 1979). However, the three persons differ in a number of important characteristics. First, the slope of the PCC's at the estimated ability level, which is the same for all three persons, shows that Person A was measured more precisely than persons B and C. An individual's guessing behavior or luck in guessing may also be inferred from the proportion of correct responses obtained on items that are well above the individual's ability level. In this case Person A does not appear to have been guessing. Similarly, carelessness can be inferred by looking at the proportion of easy items answered incorrectly by an examinee. In this case it appears that Person A was less careless than Person C. Finally, the PCC can indicate if a person is deviating from a unidimensional response pattern. Person B is answering or missing some items that are below or beyond his/her ability level.

Figure 1
Observed Person Response Curves for Three Hypothetical Persons with the Same Ability Level (θ=0.0)



From: Person Response Curve: The Fit of Individuals to the Item Characteristic Curve Models by T.E. Trabin and D.J. Weiss, 1979. Research Report 79-7. Minneapolis: University of Minnesota, Psychometric Methods Program. Reprinted by Permission.

The direct plotting of PCC's can be used to provide an indication of the unusualness or variability of an individual's response pattern on a multiple choice test. However, the direct plotting of PCC's is cumbersome and does not easily provide a quantitative indication of the unusualness of an individual's response pattern. For this reason, more useful summary statistics for quantifying and comparing the variability of individuals have been developed. Throughout this document these statistics are termed "appropriateness indices".

The use of appropriateness indices is of concern to test developers and practitioners because of the large number of reasons for inappropriate test scores. Wright (1977) and Wright and Stone (1979) define three typical reasons for response patterns to appear unusual. A "sleeping" response pattern can be produced by a person who gets bored and does poorly on the later items of the test. A "fumbling" response pattern could be produced by someone who can answer difficult items, but misses items at the beginning of the test because of confusion about the test format. A "plodding" response pattern is produced by individuals who respond to the test so deliberately and slowly that they do not have time to respond to all of the items and at some point in the test they abruptly stop getting the items correct.

Depending on the circumstances and the type of test, other reasons for a person's pattern of responses to be unusual have been cited. Wright (1977) cites item bias. For example, for certain individuals the reading level required for a mathematics test may bias certain items, or the use of special vocabulary in a reading test may be a source of bias for some of the items. He also suggests that an individual may give inappropriate responses because of being distracted, out of practice, rushed, bored, coached, or because of cheating.

Hulin, Drasgow, and Parsons (1983) suggest that a high ability person who was excessively creative might reinterpret some easy items which appeared too trivial to the examinee. They also suggest that unusual response patterns could result when an individual commits an error in filling out the answer

sheet by omitting an item and subsequently fills in the rest of the responses in the proper sequence, but incorrectly. Sato (1975) and Harnisch and Linn (1981) cite lack of preparation or sporadic study habits as causes of unusual response patterns. Levine and Drasgow (1982) mention coaching as a cause of unusual response patterns. Harnisch and Levine (1986) suggest that a person could be exposed to the subject matter in a different manner than the majority of the norm group. Birenbaum and Tatsuoka (1982) cite the example of a person consistently applying inappropriate solution strategies to certain of the questions in the test. The response pattern for such a person would depend on the test material and the organization of the test.

It is apparent from this long list of reasons that it would be very useful if persons with inappropriate test scores could be identified so that proper measures could be taken to interpret their scores and evaluate these individuals correctly. However, to use these indices effectively, test practioners must know whether certain test characteristics affect the detection of aberrance. They must know which indices are most effective for the type of aberrance that they suspect has occurred. They must decide which IRT model, if any, to choose and how many examinees are necessary for the item parameter estimation. They must decide how to set cutoff values to use for deciding that an examinee's test results are inappropriate. The purpose of this study is to provide a better understanding of these issues.

In the next chapter of this paper the major appropriateness indices that have been developed to date are described. In Chapter 3, the research that has been conducted on appropriateness indices is reviewed and the research problems are identified. In Chapter 4, the research methodology is described. In Chapter 5 the results of the research are described and discussed. Finally, in Chapter 6 the results of the study are summarized and conclusions are drawn.

CHAPTER II

DESCRIPTION OF APPROPRIATENESS INDICES

Tatsuoka and Linn (1983) distinguish between two general classes of appropriateness indices: those that are based on IRT and those that use summary statistics based on observed item responses. Hulin, Drasgow, and Parsons (1983) have used a similar classification system with two general categories. One category is termed theory-based appropriateness indices. Since, so far, IRT is the only theory that has been used for appropriateness indices, this category corresponds to the IRT class of appropriateness indices used by Tatsuoka and Linn. The second category used by Hulin, Drasgow, and Parsons is termed "heuristic" and comprises all of the indices which are not based on IRT. For convenience, in this paper a distinction is made between two general categories of indices. Those that make use of IRT have been labelled IRT-based indices and all other indices are termed heuristic appropriateness indices.

The indices which are described in this chapter were chosen for the most part because they have been the most useful or because they show some development in the evolution of appropriateness indices. While the main purpose of this chapter is to describe these indices, brief comments regarding some important characteristics of the indices are also made.

One characteristic which is discussed is the standardization of the indices. Most indices, in their original form, demonstrate a linear and/or non-linear relationship to the total score on the test and, when possible, they have been standardized. With well standardized indices it is possible for a user to interpret the index value in the same way on any test and at any ability level. Furthermore, if the distribution of the index values is known, cutoff scores at various false positive rates can be established for all tests.

Some comments are made regarding the relationship of the index to other indices, the effectiveness of the index, the cost to calculate the index, and the ease of use of the index.

Heuristic Methods of Identifying Unusual Response Patterns

All of the heuristic indices share one basic characteristic. The computation requires only traditional item statistics such as the item difficulty parameter. As a result, when contrasted with the IRT-based indices, the heuristic indices show one major limitation. The values of the indices are calculated relative to a group, which means that they can only be interpreted in relationship to that group. However, the heuristic indices have one important advantage over the IRT-based indices. Because the heuristic indices are not based on the strong assumptions of the IRT-based indices, they are likely to be applicable to a much broader range of tests and situations. In addition, the calculation of the heuristic indices may require fewer examinees and items than are necessary for the IRT-based indices.

Seven heuristic indices are described below. The first five make use of dichotomous test data (i.e. correct or incorrect answers only). These indices are presented because they are the most commonly used and researched of the heuristic indices. Although a number of other heuristic indices have been developed, Harnisch and Linn (1981) have noted that there are close algebraic links among many of these indices and a strong tendency for them to be intercorrelated. For this reason it is not necessary to examine all of the indices.

The last two of the heuristic indices that are described in this paper are rather unique and have not been researched extensively. The first, the IOV index, is unique in that it is the only polychotomous heuristic index to have been researched to date. The second, the ICI index, is also distinctive among the heuristic indices in that it is used to identify individuals who have inconsistent behaviors in approaching specific test tasks. This latter

index is highly restricted in its use because it requires the creation of a test with parallel sets of items measuring the same task.

Sato's Caution Index and the Modified Caution Index. The most well known and most researched of the heuristic appropriateness indices is the Caution Index (C_i) (Sato, 1975). With Sato's Caution Index an examinee's response pattern is compared with that of an ideal response pattern, based on the concept of the "Student-Problem Curve". The C_i is obtained by using test information arranged in a table of binary scores referred to as a Student-Problem (S-P) Table. The S-P Table is a matrix of zeros and ones, as shown in Table 1, where correct responses to items on a test are represented by 1's and

Table 1
Hypothetical S-P Table for 15 Students and 10 Items

	Item j										Examinee
Examinee i	1	2	3	4	5	6	7	8	9	0	Total n
1	1	1	1	1	1	1	1	1	1	1	10
2	1	1	1	1	1	1	1	1	1	0	9
3	1	1	1	0	1	1	1	1	0	1	8
4	1	1	0	1	1	1	1	1	0	0	7
5	1	1	1	1	1	1	0	0	1	0	7
6	1	1	1	1	0	0	1	0	1	0	6
7	1	1	0	1	0	0	1	0	0	1	5
8	1	1	1	0	1	0	0	0	1	0	5
9	1	0	1	0	0	0	1	1	1	0	5
10	1	1	1	1	0	1	0	0	0	0	5
11	1	0	0	0	0	1	0	1	0	1	4
12	1	0	0	1	1	1	0	0	0	0	4
13	1	0	1	0	1	0	0	0	0	0	3
14	1	1	0	0	0	0	0	0	0	0	2
15	0	1	0	0	0	0	0	0	0	0	1
Item Total	14	11	9	8	8	8	7	6	6	4	

incorrect responses are represented by 0's. The rows in the table represent the examinees, arranged in descending order of total test score; the columns represent the items arranged in ascending order of difficulty from left to right. For each examinee, a stair-step line, the student curve (S-Curve), is obtained by drawing a vertical line to the right of the column corresponding to the total number of correct answers obtained by the examinee on the test. For example, examinee 3 obtained 8 correct answers, so a vertical line is drawn after column number 8. The vertical lines are connected by horizontal lines to obtain the S-Curve (shown as a single line). A P-Curve (double line) can also be obtained in an analogous fashion by tabulating the number of examinees who answered the item correctly. If a test formed a perfect Guttman Scale (Guttman, 1941), the S and P curves would coincide. If the S-Curve is kept unchanged, but all of the 0's to the left of the S-Curve are changed to 1's, while all of the 1's to the right of the S-Curve are changed to 0's, we obtain the perfect S-Curve shown in Table 2.

Sato's Caution index, C_i, provides an indication of the similarity of an examinee's response pattern to a perfect Guttman vector for the S-Curve. It is defined as the complement of the ratio of the covariance of an examinee's actual response pattern and the group pattern to the covariance of the examinee's expected response pattern and the group response pattern.

$$C_i = 1 - \frac{\mathrm{cov}(\underline{y}_i, \underline{y}_\cdot)}{\mathrm{cov}(\underline{x}_i, \underline{y}_\cdot)}$$

where $\underline{y}_\cdot = (y1, y2, \ldots yn)$: The vector of the column totals

$\underline{y}_i = (y_i1, y_i2, \ldots y_in)$: The response vector for person i.

$\underline{x}_i$ = The Guttman vector of expected item responses for person i.

Table 2

Perfect S-Curve for Hypothetical Scores in Table 1

Examinee i	Item j										Examinee Total n
	1	2	3	4	5	6	7	8	9	0	
1	1	1	1	1	1	1	1	1	1	1	10
2	1	1	1	1	1	1	1	1	1	0	9
3	1	1	1	1	1	1	1	1	0	0	8
4	1	1	1	1	1	1	1	0	0	0	7
5	1	1	1	1	1	1	1	0	0	0	7
6	1	1	1	1	1	1	0	0	0	0	6
7	1	1	1	1	1	0	0	0	0	0	5
8	1	1	1	1	1	0	0	0	0	0	5
9	1	1	1	1	1	0	0	0	0	0	5
10	1	1	1	1	1	0	0	0	0	0	5
11	1	1	1	1	0	0	0	0	0	0	4
12	1	1	1	1	0	0	0	0	0	0	4
13	1	1	1	0	0	0	0	0	0	0	3
14	1	1	0	0	0	0	0	0	0	0	2
15	1	0	0	0	0	0	0	0	0	0	1
Item Total	15	14	13	12	10	6	5	3	2	1	

The term Caution Index is based on the notion that a large value for the index indicates that the observed and Guttman-scaled vectors are very dissimilar, suggesting that the examinee has an atypical response pattern; thus, some caution is needed in interpreting the examinee's score on the test.

Harnisch and Linn (1981) have suggested a modification to C_i, which yields a lower bound of 0 and an upper bound of 1. The advantage of this Modified Caution Index (C^*_i) is that it has a known upper and lower limit. Harnisch and Linn (1981) obtained a correlation of .99 for the two indices.

The C_i and C^*_i have both been found to have a fairly strong linear relationship to total test score (for C_i, r = -.17 to -.42; for C^*_i, r = -.02

to -.21) (Drasgow, Levine, & McLaughlin, 1987; Harnisch & Linn, 1981; Oltman, 1985). In one comparison to several IRT-based indices, the C_i was found to be relatively weak in detecting aberrant response patterns (Drasgow, Levine, Williams, McLaughlin, & Candell, 1987). Gafni (1987) has suggested that the C_i may be less effective than IRT-based indices because it is more strongly related to the total test score.

Personal biserial and personal point-biserial correlation. Two appropriateness indices, the personal biserial (p_{bis}) and the personal point-biserial (p_{pbis}) are conceptually very similar to the biserial and point-biserial correlations which are often used as indices of item discrimination. The personal point biserial is defined as the product moment correlation between item scores for individuals and the item difficulty (Harnisch & Linn, 1981). Donlon and Fisher (1968) proposed the personal biserial correlation based on the same two assumptions as are made in calculating the biserial correlation. The first assumption is that an individual's observed response pattern is representative of a latent variable which is normally distributed across the items of the test. The second assumption is that there is a linear regression of the observed difficulties for an individual onto the difficulties of a particular norm group.

The p_{bis} and the p_{pbis} indices have not been studied extensively. Fisher (1968) found that the personal biserial followed a fairly normal distribution. However, in several related studies Harnisch and Linn (1981) found both indices to be strongly related to total test score. The p_{bis} had a stronger linear relationship to total test score ($r = .28$ to $.63$). The p_{pbis}, had a strong curvilinear relationship ($R^2 = .41$ to $.55$). Only one study has involved a comparison of the effectiveness of these indices to other indices (Drasgow, 1982). He found the p_{bis} to be very weak in comparison to some IRT indices in detecting aberrant response patterns.

Van der Flier's U'. Van der Flier (1977) developed a very simple index which is similar in principle to Sato's Caution Index. To calculate this index, the results on the test are first ordered in the same way as in the S-P

Table. The U' index is calculated by adding the number of 1's to the right of every 0 (this value is called U) and dividing by the maximum number of 1's possible. The resulting U' index can vary from 0 to +1, with zero indicating perfect agreement with the norm group ordering of the item difficulty. Table 3 presents the results of a hypothetical 4 item test, for 6 persons. Each person obtained a score of 2 on the test.

Van der Flier's U' has been found to have a high correlation with total test score ($r = -.30$ and $r = -.54$ on two separate tests) (Harnisch & Linn, 1981).

Table 3

U and U' Values for Hypothetical Score Patterns on a Test Yielding a Total Score of Two

	Item Number				Index	
	p=.8	.7	.6	.3		
Person	1	2	3	4	U	U'
1	1	1	0	0	0	0
2	1	0	1	0	1	1/4
3	1	0	0	1	2	2/4
4	0	1	1	0	2	2/4
5	0	1	0	1	3	3/4
6	0	0	1	1	4	1

The IOV index. The Item Option Variance (IOV) index is an attractive heuristic index because it is easy to understand and compute, and because it takes into account the response alternatives that the examinees choose. For each subset of examinees in the test sample who selected choice c to item i the mean number-correct score X_{ic} is calculated. The item-option variance IOV = $Var(X_{ic})$ is considered to provide a measure of the consistency of an examinee's option choices. For individuals with inappropriate response

patterns, there should be a great deal of variance in the X_{ic} for the options selected.

The IOV has only been evaluated in one study in which the effectiveness was not assessed because the index was found to be highly related to total test score (Drasgow, Levine, & McLaughlin, 1987).

The Individual Consistency Index. The appropriateness indices described so far involve a comparison of the response pattern of an individual to that of a group norm (to the "typical" response pattern of the group). However, this typical response pattern is influenced by the ability level and learning background of the group taking the test. Thus, any of the person-fit measures that have been described so far are unstable, dependent upon the group on which the test was normed.

The Individual Consistency Index (ICI) proposed by Tatsuoka and Tatsuoka (1983) requires the individual to respond to parallel sets of items measuring the same task. The parallel sets of items are then ordered in terms of difficulty based on the individual's number of correct responses to each parallel set and the index is calculated similarly to U'.

The ICI has a very specific use in identifying individuals who have inconsistent behaviors in approaching specific tasks. It has been used by some researchers in conjunction with computerized adaptive testing and techniques designed to diagnose rule errors in completing certain tasks (Tatsuoka & Tatsuoka, 1985). However, so far its use has been limited to these circumstances and it requires the creation of parallel subtests.

In sum, all of the heuristic indices described above, with the exception of the ICI, are conceptually similar. They involve a comparison of the response patterns of an individual to that of a group. For practical purposes these indices have two disadvantages. First, the values of these indices tend to be related to the test score. Second, because they are based on classical item statistics, the values of the indices are influenced by the group taking the test. The ICI is designed for identifying individuals who have inconsistent behaviors in approaching specific tasks. It requires the

administration of parallel sets of items and as such cannot be used when only one test form is available.

Item Response Theory Based Methods of Identifying Unusual Response Patterns

In contrast to the heuristic approaches to the detection of unusual response patterns which compare the individual's response vector to that of a group norm, the second general approach to the detection of unusual response patterns is based on the IRT model of test measurement. The most commonly used item response theory models are dependent on three strong assumptions: the test is unidimensional, the items are locally independent, and the item characteristic curve (ICC) has a certain shape (an ICC plots the probability of responding correctly to an item as a function to the latent trait underlying performance on the items of the test). Appropriateness indices based on an IRT model are likely to be useful only to the extent that the test meets the assumptions of the model. An added concern with the IRT indices is that a large number of examinees may be necessary to obtain adequate estimates of the item parameters for the test.

In the standard model of IRT, if an examinee with ability Θ responds to n items on a unidimensional test, the probability of the response vector $U = \langle U_1, U_2, \ldots\ldots, U_n \rangle$ is given by

$$\text{Prob}\{U|\Theta\} = \prod P_i(\Theta)^{u_i}[1-P_i(\Theta)]^{1-u_i}$$

where $U_i = 1$ for a correct response and 0 for an incorrect response and P_i is the conditional probability of an examinee, with ability Θ, correctly answering the $_i$th item.

The most commonly used IRT model in the research on appropriateness indices is the three parameter model.

$$P_i(\Theta) = c_i + \frac{(1-c_i)\, e^{Da(\Theta-b_i)}}{1 + e^{Da(\Theta-b_i)}}$$

where

a_i is the item discrimination power,

b_i is the item difficulty,

c_i is the lower asymptote of the item characteristic curve(ICC), and

D is a scaling constant usually set equal to 1.7

However, $P_i(\theta)$ can also be calculated using the one parameter (1PL), or two parameter (2PL) IRT models. The equation for the 2PL model is the same as for the 3PL model, except that the c_i parameter is not present in the equation, while the equation for the 1PL model does not contain c_i and the a_i parameter is set to a constant.

The IRT model has been used in five types of appropriateness indices.

The l_o type of appropriateness index

The l_o index. The development of the l_o type of appropriateness index has followed an evolution designed to improve upon the original index which was developed by Levine and Rubin (1979). The original l_o index is based on the probability of a pattern of responses U. A relatively low maximum of the function of Prob(U|θ), is considered to be indicative of an unusual response pattern which led Levine and Drasgow (1982) to define l_o as max Prob (U|θ).

Unfortunately, the l_o index results in smaller values for individuals who answer more items and is thus inappropriate in circumstances where there are omitted items. When omitted items are present in the response vector the l_o index is only computed for the answered items. The l_o index has also been found to be related to total test score (e.g. Birenbaum, 1985).

The l_z index. To improve the l_o index Drasgow (1982) proposed dividing l_o by the number of items answered by the examinee. The l_z index, $l_z = \exp(l_o/n)$, where n is the number of items attempted by the examinee, can be interpreted as the geometric mean response probability.

While this index is expected to be less sensitive to the effect of test length than the l_o index, it still suffers from being dependent on the total test score (Drasgow, 1982).

The l_h index. Drasgow, Levine, and Williams (1985) introduced the use of a polychotomous model for appropriateness measurement. The model is based on the same assumptions as the standard three parameter logistic model regarding unidimensionality and local independence, but it extends the most common form of the model, the dichotomous form, where only the correct option is considered, to all of the option choices including an omit. As for the l_o

and l_g indices, the l_h index has also been found to be dependent on the total test score (Drasgow, Levine, & Williams 1985).

The Lz and zh indices. While the problem caused by omitted items was solved with the l_g and l_h indices, an additional improvement in the indices was seen with the standardization of l_o and l_h. Drasgow, Levine, and Williams (1985) were able to estimate the conditional means and standard deviation for the indices. They thus defined the standardized l_o as follows:

$$Lz = \frac{l_o - E(l_o)}{[var(l_o)]^{1/2}}$$

Similarly, Drasgow, Levine, and Williams were able to define a standardized polychotomous index zh:

$$zh = \frac{l_h - E(l_h)}{[var(l_h)]^{1/2}}$$

The 3PLM versions of the Lz and zh have both been found to be fairly well standardized (Drasgow, Levine, & Williams, 1985; Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, McLaughlin, & Candell 1987; Gafni, 1987). The Lz has been found quite effective in comparison to other indices (Drasgow, Levine, & Williams, 1985; Drasgow, Levine, & McLaughlin, 1987; Gafni, 1987). The zh has been found effective in detecting individuals who have omitted a large number of responses (Drasgow, Levine, & Williams, 1985), but less effective than the Lz in detecting other types of aberrance (Drasgow, Levine, McLaughlin, & Candell, 1987).

In summary, the standardized forms of the l_o type of indices have proven to be among the most effective of the appropriateness indices reported in the literature. The Lz index not only eliminates most of the dependence of the index on the ability level of the individual, but also allows meaningful comparisons of individuals who have omitted a different number of items on the test. The zh also seems to be fairly well standardized and may be particularly powerful in detecting examinees who have omitted a large number of items.

Appropriateness indices based on the Gaussian model

Levine and Rubin (1979) developed two appropriateness indices based on what they termed the "Gaussian model" of item response theory. Unlike the standard IRT model, in the Gaussian IRT model the latent trait, Θ, is not assumed to be unidimensional, and the values of the Θ's are allowed to vary for each response. A new ability level, Θ_i, is considered to have been sampled for each item. These Θ's are assumed to be independently sampled from a normal distribution having a mean of Θ_o and a variance of σ_g^2.

The LR index . The first index is based on the principle that the response pattern of an individual who responds in an aberrant fashion to a test will be better fitted by the Gaussian model than by the standard IRT model, for which it is assumed that the individual has one level of Θ throughout the test. To arrive at this index, Levine and Rubin (1979) first defined an index, l_a, in a similar fashion to l_o.

$$l_a = \log \max \text{Prob}(U|\Theta_i, \sigma_g^2)$$

Because the standard IRT model is a submodel of the Gaussian model, the likelihood ratio (LR) test was proposed as an index which would indicate the improvement in the fit of the model that is obtained by allowing the Θ's to vary during the test. The LR index is simply the difference between l_o and l_a.

$$LR = l_a - l_o$$

The σ_g^2 index. The second index based on the Gaussian model is the variance, σ_g^2, of the Θ_i distribution. This index should be zero when a vector of responses is determined by a single ability. The extent to which σ_g^2 is greater than zero indicates the degree to which the ability varied in the model and therefore the inappropriateness of the response pattern.fitting

The appropriateness indices based on the Gaussian model have usually given similar results to the l_o type of indices. Since the Gaussian indices require more computing time, they have been used infrequently.

Fit Statistics

A number of appropriateness indices which are based on the IRT model have been derived from the work of Wright (1977) with the Rasch model and are based on the fit of an individual's response pattern to the pattern which is expected, based on the probabilities of correct responses generated by the model. The deviation of any observed response from the actual response is termed the residual for item i (R_i).

$$R_i = u_i - P_i(\theta)$$

where u_i = 1 or 0 and;

$P_i(\theta)$ is the probability of a correct response according to the model.

Two statistics based on the concept of the standardized residual are most commonly associated with the Rasch model. The first, the unweighted total mean square statistic (U1), is simply the sum of the standardized residuals divided by n, the total number of items in the test.

$$U1_j = 1/n \sum_{i=1}^{n} [(u_{ij} - P_{ij})^2/P_{ij}(1 - P_{ij})]$$

A second statistic often referred to in the literature, is the total weighted mean square statistic, W1, which is a weighted version of U1

$$W1_j = \sum_{i=1}^{n} (u_{ij} - P_{ij})^2/\sum_{i=1}^{n}[P_{ij}(1 - P_{ij})]$$

where u_{ij} is the observed response on item i for person j and n is the number of items in the test.

While these mean square statistics were conceptualized within the framework of the Rasch model, their extension to the two or three parameter logistic models is simply a matter of calculating the P(θ) using the more complex models and substituting the results into the above formulae. Rudner (1983) has termed the 3 parameter fit statistics corresponding to the statistics described above as W3 and U3 respectively; Birenbaum (1985) termed the 2 parameter fit statistic corresponding to the unweighted fit statistic, U2.

While it has been hypothesized that these mean square fit statistics have a mean of one and a standard deviation that can be estimated by various formulae (Smith, 1982), simulations of responses to a 30 item test when the data fit the model have not supported this hypothesis (Smith, 1982). The U3 index has been found to be related to total test score (Drasgow, Levine, & Mclaughlin, 1987; Gafni, 1987). The W3 index appears to be better standardized than U3 (Drasgow, Levine, & McLaughlin, 1987), but not as well standardized as ECIZ2[1], ECIZ4[1], and Lz (Drasgow, Levine, & McLaughlin, 1987; Gafni, 1987; Noonan, 1989).

When compared to standardized IRT indices, the effectiveness of the fit statistics has not been consistent. In one study, Drasgow, Levine, McLaughlin, and Candell (1987) found the U3 to be among the most effective of the indices, but in other studies it was found to be less effective (Drasgow, Levine, & McLaughlin, 1987; Gafni, 1987). Similarly, the W3 was among the most effective of the indices in one study (Drasgow, Levine, & McLaughlin, 1987), but not in others (Drasgow, Levine, McLaughlin, & Candell, 1987; Gafni, 1987; Noonan, 1989).

Smith (1982, 1985) presented an additional appropriateness statistic based on the 1PLM. The unweighted between statistic is based on the postulate that if the data fit the model, the overall ability estimate should accurately predict the person's score on any subset of items. By comparing the person's predicted score ($\sum E_i$) with the observed score ($\sum O_i$) on any subset of items it is possible to test the fit of the data to the model. The unweighted between statistic (UB) is

$$UB = (1/J-1) \sum_{j=1}^{J} (\sum_{L=1}^{L_j} Oi - \sum E_i)^2 \; / \; \sum [E_i(1-E_i)]$$

where J is the number of subsets and L_j is the number of items in each subset.

---

[1]These two indices are discussed below.

Gafni (1987) extended the UB to the 3PLM. The UB index has been found to be relatively effective in detecting aberrance that may have been caused by differential familiarity or ability on certain specified items within the test and in detecting aberrance caused by guessing on the last items of certain tests (Smith, 1985; Gafni, 1987).

The distributions of the fit statistics have not been examined extensively, but they are thought to be influenced by each combination of test length, distribution of item difficulty, and person ability (Smith, 1982). Research has shown them to be quite related to total test score (Drasgow, Levine, & McLaughlin, 1987).

The Extended Caution Indices

The Extended Caution Indices (ECI) have been grouped with the IRT indices because they make use of IRT terms. The ECI's originate in the Caution Index, which was developed based on 0-1 scoring, and extend the theory by using probabilities based on IRT theory (Tatsuoka, 1984; Tatsuoka & Linn, 1983).

As shown above, the formula for C_i is

$$C_i = 1 - \frac{\mathrm{cov}(\underline{y}_i, \underline{y}_\cdot)}{\mathrm{cov}(\underline{x}_i, \underline{y}_\cdot)}$$

where $\underline{y}_\cdot = (y1, y2, \ldots yn)$: The vector of the column totals

$\underline{y}_i = (y_i1, y_i2, \ldots y_in)$: The response vector for person i .

$\underline{x}_i$ = The Guttman vector of expected item responses for person i.

The first extended caution index, ECI1, is based on substituting P_i for x_i in the formula for C_i.

$$ECI1 = 1 - \frac{cov(\underline{Y}_i, \underline{Y}.)}{cov(\underline{P}_i, \underline{Y}.)}$$

where $\underline{P}_i = (P_{i1}, \ldots, P_{in})$ for n items and;

P_{ij} is the probability associated with examinee i answering item j correctly.

The second ECI makes use of the conceptual similarity between G, the group response curve of IRT, and Y.. The group response curve (G) is an average function of N person response curves. The ECI2 goes one step further than the ECI1 in the formula for C_i by substituting the group response curve vector, here called ($\underline{G}$), for ($\underline{Y}.$).

$$ECI2 = 1 - \frac{cov(\underline{Y}_i, \underline{G})}{cov(\underline{P}_i, \underline{G})}$$

Tatsuoka and Linn (1983) have also derived ECI3, which uses the same terms as ECI2 but is given by the complement of a correlation ratio:

$$ECI3 = 1 - \frac{corr(\underline{Y}_i, \underline{G})}{corr(\underline{P}_i, \underline{G})}$$

The actual difference between ECI2 and ECI4 lies in their numerators. ECI4 indicates the extent to which the observed response vector, $\underline{Y}_i$, covaries with the probability of correct item response vector $\underline{P}_i$:

$$ECI4 = 1 - \frac{cov(\underline{Y}_i, \underline{P}_i)}{cov(\underline{G}, \underline{P}_i)}$$

Tatsuoka and Linn (1983) have indicated that conceptually ECI4 is the probabilistic extension of the ICI index, showing how similar an individual's response patterns is to the theoretical person response curve at the individual's θ.

Tatsuoka and Linn (1983) also defined ECI5 and ECI6, but these two indices have never been researched. Similarly, ECI3 has not been researched.

The ECIs give inflated values at both extremely high and low total scores. In order to standardize these indices, Tatsuoka and Tatsuoka (1982) derived conditional expectations and variances for all of the ECI's. The standardized forms of the ECI's have a Z following the I (e.g. the standardized ECI4 is ECIZ4).

Most of the research on the extended caution indices has involved the ECIZ2 and the ECIZ4. In all of the studies in which one or both of these indices have been compared to other indices they have always been among the most well standardized (Harnisch & Tatsuoka, 1983; Drasgow, Levine, & McLaughlin, 1987; Gafni, 1987; Noonan, 1989). They have also been among the most effective at detecting aberrance (Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, McLaughlin, & Candell, 1987; Noonan, 1989).

Optimal Appropriateness Indices

Levine and Drasgow (1984; 1988) introduced a general method for determining the maximum rate of detection of a specified form of aberrance that can be achieved by any index based on the IRT model. These indices require the probabilities of observing the response vector first by assuming that it was generated by a non-aberrant process and then by assuming that it was generated by a specified aberrant process. Two optimal indices have been reported in the literature. The polychotomous model optimal index, LRh, requires the probabilities of observing the polychotomous response vectors, while the dichotomous model optimal index, LR3, is based on dichotomously scored item responses. Technical details about these indices and an efficient computing algorithm are given by Levine and Drasgow (1984; 1988).

So far the principal purpose of these "optimal" indices has been to provide benchmarks for evaluating practical indices, by indicating the highest detection rates possible under a given set of restrictions. If a practical index is nearly as powerful as the optimal index then there is no point in trying to improve upon the practical index. This type of information can

point the way to more powerful indices. It is possible, for example, to compare the optimal power of a polychotomous index to that of a dichotomous index. If the optimal polychotomous index is significantly more powerful than the dichotomous index, it may be worthwhile trying to develop practical polychotomous indices.

While the optimal indices are useful as benchmarks, they are not well suited for most practical situations. For each application of the optimal indices the amount and type of aberrance must be specified (e.g. 10% high or 10% low). These limitations only allow the optimal indices to be used for a specific hypothesis regarding the level and type of aberrance present in the sample. Another disadvantage of these indices is that they have not been standardized.

Summary

There are two general classes of appropriateness indices: heuristic indices and IRT-based indices. Heuristic indices make use of traditional item statistics, such as item difficulty, and are based on the principle that normal response patterns are characterized by the tendency of the examinee to correctly answer items that are easier for the group, while making mistakes on items that are more difficult for the group. Many of these indices are mathematically related and are highly correlated. The most researched of the heuristic indices is the Modified Caution Index, C^*_1.

The heuristic indices tend to be related to total test score, which means that the same values of the indices have different meanings for individuals depending on their scores on the test. None of the heuristic indices has been standardized. In comparison to the IRT indices, the heuristic indices have not been as effective in detecting aberrance.

Only one polychotomous heuristic index, the IOV, has been reported. It was found to be highly related to the individual's total test scores and was considered to be ineffective in detecting aberrant response patterns.

The IRT-based indices make use of item response theory and as such should only be used when the assumptions of IRT have been met. As with heuristic indices, the IRT-based indices are related to the total test score. For some of these indices standardized forms have been developed, using estimates of their conditional means and standard deviations. Five types of appropriateness indices have used the IRT model.

The first type of IRT index, the l_o type, is based on the likelihood of a response pattern according to the IRT model that is being used. The less likely the pattern, the higher the absolute value of the index. The standardized version of this index, Lz, has been among the most successful of this type of index in detecting aberrant response patterns.

The second type of IRT index, the Gaussian type, is similar to the l_o type, but is based on the "Gaussian model" of IRT in which the values of θ are allowed to vary with each response. This type of index has given similar results to the l_o type and has not been used frequently, primarily because it is more expensive to compute.

A third type of IRT index, the fit statistic, is based on the deviation of an observed response from the response that is expected according to the IRT model. The two fit statistics which are most commonly used are the total weighted and total unweighted mean square fit statistics, W, and U respectively. While these indices have sometimes proven relatively effective in detecting inappropriate response patterns, their performance has been inconsistent. A third fit statistic, the unweighted between fit statistic, UB, compares the observed responses to those expected based on the person's ability as indicated by the entire test. The UB index has been found effective for detecting aberrance caused by differential familiarity with certain items on a test and with detecting aberrance caused by guessing on the last items of a test.

A fourth type of IRT index, the extended caution index (ECI), is based on the formula for the Caution Index, C_i, and substitutes various terms derived from IRT. While there are six ECIs, the two most researched ECIs are

the standardized versions of ECI2 and ECI4. ECI2 is based on the relationship between the individual's observed response pattern and the group probabilistic pattern. ECI4 is a measure of the similarity of the individual's response pattern to the theoretical person response curve at the individual's θ. The standardized versions of these two indices, ECIZ2 and ECIZ4, have been among the most effective in detecting inappropriate response patterns.

A fifth type of IRT index, the optimal appropriateness index, has been developed primarily for assessing the maximum rate of detection that is possible for a clearly specified form of aberrance. Because the index requires that the aberrance be specified in detail, it is not useful under practical circumstances where these details cannot be known.

CHAPTER III

RESEARCH ON THE EFFECTIVENESS OF APPROPRIATENESS INDICES

Two general classes of study have been used to determine the effectiveness of appropriateness indices. The first approach is to ascertain if the test results of examinees who are identified by the indices as having inappropriate response patterns are related to other criteria that they are expected to predict. For example, Chatman and Nelson (1984) anticipated that students who reported that they had guessed frequently on a test would have more inappropriate response patterns than those who guessed less frequently. Studies using the first approach are unreliable for examining the properties of appropriateness indices for three reasons. One reason is that in all of these studies the amount of aberrance is unknown. A second problem associated with some of these studies is that they are based on a contentious relationship of aberrance to the criterion variable. If the predicted relationship is not found some might argue that it is because no such relationship should exist. A third serious limitation of these studies is that they provide very little information regarding the comparative effectiveness of different indices, simply because in these studies only one or two indices were used.

In order to overcome the limitations of the studies involving expected relationship, researchers have turned to a second approach for determining the effectiveness of appropriateness indices. In these studies there is no question that the simulees actually do have aberrant response patterns, and it is possible to quantify the amount of aberrance.

Before reviewing this research, it is useful to understand the different procedures that have been used for simulating aberrance. The procedure that the majority of researchers have used will be referred to as the Levine procedure simply because it was first used by Levine and Rubin (1979) in one

of the first simulation studies on aberrance. It is described in some detail by Rulin, Drasgow, and Parsons (1983, p.131).

In the first step of this procedure item parameters are obtained which are used to estimate ability and compute appropriateness indices throughout the study. Usually, a test norming sample consisting of N examinee responses to n items is selected for the estimation of the item parameters. In the second step a non-aberrant sample of response vectors is generated using the item parameters obtained in the first step and a specified set of abilities for the simulated examinees. In the third step examinees are simulated with either spuriously high or spuriously low scores by modifying the response vectors.[2] Examinees with spuriously high scores are simulated by randomly selecting k% of the examinees' responses without replacement and recording these responses as correct regardless of the original response. Examinees with spuriously low scores are simulated by randomly selecting k% of the examinees' responses without replacement and replacing the response by assigning a 1/c chance of obtaining the correct response, where c is the number of choices. In the fourth step appropriateness indices are computed for the non-aberrant and aberrant response vectors. The effectiveness of the index is evaluated based on the extent to which it can separate aberrant and non-aberrant response vectors solely on the basis of appropriateness index scores.

One disadvantage of the Levine technique is that the number of responses changed in a response vector is different from one simulee to another. For some of the response vectors there may be few changes and aberrance may be difficult to detect for this reason. As a result, in some studies (Levine & Rubin, 1979; Levine & Drasgow, 1982) the researchers used only those response vectors in which more than 10% of the items had been changed.

A variation of the Levine procedure which has been used in two simulation studies (Rudner, 1983; Nconan 1989) does not result in too few

---

[2]Response vectors with spuriously high scores are said to have "high aberrance, while response vectors with spuriously low scores are said to have "low aberrance".

changes in the response vectors. Instead of randomly selecting k% of the examinees' responses regardless of whether they are correct or incorrect, only incorrect responses are selected for the spuriously high modification and only correct responses are selected for the spuriously low modification. Thus, spuriously low scores are obtained by randomly selecting a specified number of correct responses and changing them to incorrect, while spuriously high scores are obtained by changing a specified number of incorrect scores to correct. This procedure, which will be called the Rudner procedure, results in a known percentage of items on the test that have been changed.

Using the Levine or Rudner procedures for generating aberrant response patterns, items from throughout the test are selected randomly. These procedures represent some reasons for aberrance, such as lack of motivation, which are likely to randomly affect any item in the entire test. However, certain causes for aberrance, such as test anxiety at the beginning of a test, affect only some parts of the test. Some researchers have simulated these causes for aberrance, sometimes using the Levine procedure, but only applying it to specific parts of the test. Gafni (1989) has identified these as studies of "systematic aberrance" and has termed studies in which items from throughout the test are affected as studies of "random aberrance". Some studies have involved a mix of both procedures. In this paper, the studies that deal only with random aberrance are reviewed first followed by those that involve systematic aberrance.

The first studies in which random aberrance was simulated predated the introduction of the standardized indices. As mentioned above, unstandardized indices are not very useful in practical applications because they are related to ability level as measured by the test. However, several of these studies are presented because they demonstrate the potential of the indices and represent developments in the methodology used for simulations.

Levine and Rubin (1979) investigated the effectiveness of l_o, LR, and σ^2_g indices for detecting unusual response patterns using data generated according to the 3PL model and item parameters based on the 85 item Scholastic Aptitude Test, Verbal Section (SAT-V). For both spuriously high and spuriously

low aberrance, the Levine technique was used to generate aberrance at the 4%, 10%, 20%, and 40% levels.

In this study, as well as in a number of subsequent studies, Receiver Operating Characteristic (ROC) curves were used to evaluate the effectiveness of the indices. ROCs are created by first calculating the index values for each response vector in both the non-aberrant and aberrant sample. A large number of criterion values are chosen as possible cut-off points for deciding if a response vector is aberrant. For each of these values the proportion of the aberrant sample correctly identified as aberrant are calculated (this is termed the "hit rate") and the proportion of the non-aberrant sample incorrectly identified as aberrant is calculated (this proportion is termed the "false positive rate"). The plot of these pairs of proportions is termed the ROC curve. An effective index is one which has a good hit rate at a low false positive rate. Levine and Rubin (1979) found similar ROC curves for the three indices and reasonable detection rates were obtained. For example, at the 20% level of aberrance the LR index identified approximately 40% of the aberrants at a .05 false positive rate.

This study was seminal in that it demonstrated the possible usefulness of the three indices for detecting aberrance. However, the study was limited to unstandardized indices and was based on the simulation of response vectors with no omits. This limits the generalizability of the results. In fact the SAT-V test responses, on which the parameters of the simulation were based, contained omits.

As mentioned above, the ROC curve was used in this study and in some other studies to illustrate the effectiveness of the indices. However, the detection of aberrance has limited practical usefulness when the false positive rates are very high. For this reason some researchers have not presented the high false positive rates for the ROC curves. Others have dispensed with the use of ROC curves altogether and have simply reported the proportions of the aberrants detected for a limited number of discrete false

positive rates. In most studies these false positive rates have not exceeded .10.

Levine and Drasgow (1982) conducted two simulations to answer several questions which had been left unanswered by Levine and Rubin (1979). In both studies the l_o index was used. In the Levine and Rubin study all of the appropriateness indices had been calculated with the item parameters used in generating the simulated responses. Levine and Drasgow's first study was designed to determine whether detection rates would be similar with estimated item parameters from the SAT-V. Aberrant response patterns were simulated and index values were calculated with both LOGIST item parameter estimates and item parameters used in generating the simulated responses. There was close agreement on the values of l_o obtained for the two sets of item parameters.

In the second study 2800 nonaberrant and 200 aberrant samples (at 20% aberrance) were merged and estimated item parameters were obtained from the merged sample. The l_o scores were calculated, first based on item parameters estimated from the merged group and second using the item parameters used in generating the simulated responses. The results were very similar, suggesting that the presence of aberrants in the sample would not necessarily affect appropriateness measurement.

These studies are important in that they demonstrated that the l_o index could be robust to errors in the estimation of item parameters and to the inclusion of aberrants in the test norming sample. However, the conditions for each of the studies were very limited, raising some doubt about the generalizability of the results.

Drasgow (1982) compared the effectiveness of three IRT indices (the l_g, the LR, and the $\sigma^2_{g)}$, and one heuristic index (the p_{bis}). Two other objectives of the research were to compare the results from the IRT indices based on the 1PL model with those based on the 3PL model, and to compare the results obtained for the IRT indices based on parameters estimated from a sample size of 500 versus those obtained with a sample size of 3000.

In this study, the responses of 10,000 examinees on the 95 item GRE-V were used as the data base. The Levine procedure was used for 21% of the test items for both high and low aberrance. Item parameters were estimated twice: first based on the first 3000 examinees, and a second time based on 500 examinees by selecting every 10th examinee from the GRE-V tape. Index scores were calculated for the aberrant groups and for the first 3000 examinees in the data base.

The indices were compared at six false positive rates ranging from .01 to .25 and the following results were noted: The p_{bis} was less effective than the IRT indices. For the spuriously low examinees, the three IRT indices performed about equally well in detecting aberrant response vectors for both the item parameters estimated with 3000 and 500 examinees. For the spuriously low sample, the indices were more effective with the 3PLM than with the 1PLM. The results for the spuriously high sample were poor and were about equal for the 1PL and 3PL models.

The results of this study contradict the results of Levine and Rubin (1979), where it was found that spuriously high response vectors were more detectable than spuriously low response vectors for simulated SAT-V data. This suggests that the detectablity of different types of aberrance depends on the characteristics of the items in the test. Drasgow noted that the c parameters of the difficult items in the GRE-V test were generally higher than in the SAT-V. Consequently, a low ability examinee who correctly answered difficult items would be much more easily detected on the SAT-V than on the GRE-V.

The findings of this study raise some interesting questions. The finding that the 3PLM indices were more effective than the 1PLM, even when a sample size of 500 was used to estimate the item parameters, is important in practical applications, where sample size may be restricted. It would be important to know if this result is generalizable to other tests. It would also be important to know if these results are generalizable to standardized indices and if the same result would be obtained with other sample sizes. It

would also be useful to compare the performance of other models. In practical situations the choice may need to be made among the 1PL, 2PL, and 3PL models.

Rudner (1983) examined the effectiveness of nine appropriateness indices: U1, W1, U3, W3, l_z, P_{bis}, P_{pbis}, NCI (an index which is strongly related to U'), and C^*_i. The responses for two tests were simulated based on the item parameters of the 80 item SAT-V and a 45 item general biology examination.

The same procedure for generating aberrance was followed for both tests. For the SAT-V test, experimental groups were created by first selecting groups of 100 examinees from a normally distributed non-aberrant population with a mean ability of 0 and a standard deviation of 1. Four groups of spuriously low examinees were created by altering 5, 10, 15, or 20 items to be incorrect, while four groups of spuriously high examinees were created by altering 5, 10, 15, or 20 items to be correct. Thus, these four samples represented approximately 6%, 13%, 19%, and 25% aberrance. For the general biology examination the same procedure for generating aberrance was followed except that only three groups were formed, with 5, 10, and 15 items changed to represent approximately 11%, 22%, and 33% of the total number of items in the test. The non-aberrant group consisted of 2,000 normally distributed examinees with unaltered responses patterns. The effectiveness of the indices was compared at the .05 false positive rate.

The data for both tests showed a number of interesting results. The W1, P_{bis}, P_{pbis}, NCI, and C^*_i indices were all highly correlated (-.99 to +.96). The effectiveness of the indices increased as the amount of aberrance was increased. Generally, better detection rates were obtained for the IRT indices. Detection rates for the longer test were generally higher than for the shorter test. Overall, W3 tended to be the most powerful index. However, the test and the type of aberrance affected the power of the indices, with some indices being more powerful for some conditions, but not for other conditions. For example, the highest detection rates (50%-80%) for the longer test were obtained with the W3, U3, and l_z indices for the 13% and 19%

spuriously high scores. The highest detection rates for the shorter test (40%) was obtained with W1 for the 19% spuriously low scores.

This study was one of the first in which the comparison of several types of indices was made. Unfortunately, it did not include any standardized indices. It is also unfortunate that the indices were only compared at one false positive rate. The relative performance of the indices might have been different at other false positive rates.

In this study the technique for generating aberrance results in a known number of items that are changed in the test. This is an advantage over the Levine technique where the number of items changed for any level of aberrance is not constant, but is influenced by the examinees' original levels of ability.

Harnisch and Tatsuoka (1983) computed 14 appropriateness indices: ECI 1-5, ECIZ (1,2,4), l_o, l_n, U1, U3, W1, and Lz, based on the responses of 2,437 students who took the National Assessment of Educational Progress mathematics tests. The researchers were primarily interested in the relationship of the indices to total score, the distribution of the index scores, and the relationship of the indices to each other. While the actual effectiveness of the appropriateness indices was not investigated in this study, the study is important and has been included in this review because so many indices were examined.

When the curvilinear as well as the linear relationship to total test score was taken into account, the ECIZs were the least related to total test score with Lz the next least related. The ECIZs approximately followed a normal distribution while Lz provided the second best approximation to the normal distribution.

All of the indices were somewhat related, with most of the correlations above .70. The U1 showed the least relationship to the other indices. The ECIs and ECIZs showed the greatest relationship with each other (all correlations >.80).

Drasgow, Levine, and Williams (1985) investigated the distributions and effectiveness of the l_o, l_h, Lz, and zh indices. A sample of approximately 75,000 examinees who had taken the 85 item SAT-V served as the data pool for these studies. From this data pool 3000 examinees with an approximately normal distribution of ability were selected and item parameters were calculated.

Two studies were conducted to examine the distributions of the indices. First a group of 464 nominally non-aberrant examinees (i.e. this sample may have contained aberrants) with ability in the interval -2.05 to +2.05 was selected from the sample of 75,000 examinees. Scatterplots of the distributions of the index values for this group suggested that l_o and l_h were quite related to ability while Lz and zh were much less so.

The researchers conducted a second series of investigations of the distributions of zh and Lz involving a sample of 3478 nominally non-aberrant examinees as well as a simulated sample of 4000 examinees. Using Kolmogorov-Smirnov tests, they concluded that zh and Lz closely approximated a standard normal distribution.

To investigate the power of the two standardized indices to detect aberrant response patterns, the response patterns of 300 examinees from each of five ability groups [(-2.05 to .80); (-.80 to .24); (-.24 to .24); (.24 to .80); and (.80 to 2.05)] were modified at the 10%, 20%, and 30% rate for both spuriously high and spuriously low response patterns. The Levine technique of generating aberrant responses was used.

For this study ROC curves were plotted for false positive rates of less than or equal to .20. The zh produced higher rates of detection for the spuriously low examinees than Lz, but the converse was true for the spuriously high examinees. Intuitively it would seem that zh should be more powerful in detecting aberrant responses than Lz because it is sensitive to the pattern of incorrect responses. The researchers concluded that zh was ineffective in detecting high aberrance because most non-aberrant examinees choose a few improbable incorrect responses. This caused a great deal of "noise" for the zh index, which reduced its effectiveness. The researchers suggested that it

may be possible to develop other polychotomous indices that will not be affected in this way.

The results showed that detectability increased with an increase in the amount of simulated aberrance. The results also indicated that detection rates increased as the ability of the examinees became more extreme. For example, for the 20% low modification at the .05 false positive rate the highest detection rate was obtained in the highest ability group (approximately 63% for Lz). For the 20% high modification the highest detection rate was obtained in the lowest ability group (approximately 68% for Lz).

This study is the first in which the detection of aberrance for different ability levels was reported. The Levine technique was used for the generation of aberrance and the greater detectability of aberrance at one extreme of ability was probably caused by the higher number of responses that were changed for these ability levels. However, because the actual number of responses changed was not reported, the exact nature of this relationship is not clear. To clarify this relationship it would have been useful to report the effectiveness of the indices in terms of the number of responses changed or the change in ability level caused by the aberrance. This would also permit an evaluation of the effectiveness of the indices in practical terms.

Birenbaum (1985), used a 2PL model to examine nine IRT indices, ECI1, ECI2, ECI4, ECIZ1, ECIZ2, ECIZ4, l_o, Lz, and U2, in order to determine their relationship to total test score and their effectiveness. The study was based on 13 multiple-choice anchor items from parallel forms of a reading comprehension test of English as a Second Language. Item parameters were computed from the results of 1,864 examinees.

All of the indices were found to be highly intercorrelated (-.99 to +.97) except for l_o and U2. The indices all showed some relationship to total test score, in some cases quite strong when the curvilinear relationship was considered. Overall l_o was the most related to total test score. The relationship of the standardized indices to total test score was less than

their unstandardized counterparts. Of the standardized indices, ECIZ1 showed the strongest curvilinear relationship to total test score (R^2 = .06).

The indices were used to try to identify 72 aberrant response patterns, 42 cases which were considered inappropriate because the students had not written their name on the answer sheet, which was interpreted to mean that the students had not been motivated to respond to the test, and 30 response patterns which were randomly generated. The non-aberrant sample consisted of 238 test results where students had written their names on the answer sheet. The detection rates for the different indices were presented at different false positive rates and were therefore difficult to compare. Lz had the highest detection rate at 83%, but at a false positive rate of 20%.

While this study is somewhat useful in showing the relationship of the indices to each other and to the total test score, it is of limited usefulness in demonstrating the effectiveness of the indices. The number of items in the test is quite small for a paper and pencil test. In addition, the size of the sample of the aberrant examinees was small and the use of the criterion of not writing one's name on the answer sheet as an indication of lack of motivation, and therefore of aberrance, is tenuous.

Drasgow, Levine, and McLaughlin (1987) investigated eleven indices: Lz, C_i, ECIZ2, ECIZ4, U3, W3, IOV, LR, a jacknife variance estimate (JK), and two indices based on the "flatness" of the likelihood function, the "likelihood function curvature statistics" which will be referred to as the O/E indices. The latter two indices involve information that is mainly a function of the c parameter and they are therefore difficult to understand and interpret.

The first part of the study was designed to determine which indices were well standardized. Five samples of non-aberrant response vectors were created from the item parameters of the 85 item SAT-V based on the 3PLM. Each sample contained 200 response vectors in five ability intervals: low (-2.05 to -1.5), moderately low (-.70 to -.55), average (-.05 to .05), moderately high (.55 to .70) and high (1.49 to 2.05). For a well standardized index, the distribution of index values is nearly the same in sub-populations of non-aberrant

examinees differing only in ability. A comparison of the ROC curves for the five ability levels indicated that the IOV, C_1, and U3 were highly related to ability levels and therefore were poorly standardized. The rest of the indices were considered fairly well standardized.

To examine the effectiveness of the indices that were considered well standardized, 2000 aberrant response vectors were generated for each of 12 conditions, using the Levine procedure. The 12 conditions consisted of combinations of type of aberrance (high or low), severity of the aberrance (15% or 30%), and ability level of the aberrant sample (three levels). Three ability levels, very low (0-9th percentile), low (10th-30th percentile) and low average (31st-48th percentile) were used for the spuriously high sample, while very high (93rd percentile and above), high (65th-92nd percentile) and high average ability (49th-64th percentile) levels were used for the spuriously low samples. Index values for false positive rates ranging from .001 to .10 were established by generating 4000 non-aberrant response vectors.

The JK index and the O/E indices were not considered effective in detecting aberrant response patterns. The Lz, W3, ECIZ2, and ECIZ4 had the best overall rates of detection across different conditions. These indices were quite effective when the treatments were applied to the extreme ability levels; however, they were not very effective in the middle of the ability distribution. For example, Lz detected 87% of the aberrants at the .05 false positive rate for the 15% spuriously low treatment with very high ability response vectors, while the corresponding detection rate for the high average ability range was 35%.

As mentioned above, in simulation studies it would be useful to report the amount of simulated aberrance in terms of the number of items changed or the change in examinee's ability level. This critique applies to this study and to all of the studies with simulated aberrance which are reviewed in this paper.

Drasgow, Levine, Williams, McLaughlin, and Candell (1987) investigated the effectiveness of nine practical indices based on polychotomous and dichotomous test models: Lz; zh; W3; U3; C_i; ECIZ2; ECIZ4; JK, and O/E. Item parameters from the 30 item Armed Services Vocational Aptitude Battery Arithmetic Reasoning Test were used for simulating item responses based on the 3PLM.

The Levine technique was used to generate 2000 aberrant response vectors in each of 16 conditions. The conditions were the result of varying three factors: the type of aberrance (high or low), the severity of the aberrance (17% or 33% modification), and the ability level of the aberrant sample. Four ability levels for the spuriously high samples were labelled very low (0-9th percentiles), low (10th-30th percentiles), low average (31st-48th percentiles) and high average (49th-64th percentiles). Similarly, four ability levels for the spuriously low samples were labelled very high (93rd percentile and above), high (65th-95th percentiles), high average (49th-64th percentiles) and low average (31st-48th percentiles).

For purposes of comparison, a non-aberrant sample of 4000 response vectors was generated from a normal (0,1) ability distribution. The indices were evaluated at false positive rates from .001 to .10. Several important results were obtained.

(1) The Lz, U3, ECIZ2, and ECIZ4 were the most effective indices.

(2) The C_i, JK, and O/E indices were not very effective.

(3) The Lz was more effective than the zh for high and low aberrance.

(4) Detection rates were far better at the extremes of ability. For example, at the .05 false positive rate for the 17% spuriously high condition the detection rate was 33% for the 0%-9% ability group. For the 17% spuriously low condition detection was highest for the 93%-100% group (51% detection rate at the .05 false positive rate).

The general design of this study is similar to some of the earlier studies and the results tend to correspond. No explanation was given as to why the zh was so ineffective.

Noonan (1989) used a Monte Carlo approach to investigate the characteristics of the distributions of Lz, ECIZ4, and W3 as well as the effectiveness of these indices. To examine the distributions of the indices, 2000 non-aberrant response patterns were generated for each of four conditions of test length (40 items and 80 items) and IRT model (2PLM and 3PLM). With the use of Datagen, a program written by Hambleton and Rovinelli (1973) and modified by Carlson (1985), examinee response patterns were generated with a normal distribution (0,1) of ability. Datagen was also used to generate a uniform distribution of item difficulty and item discrimination parameters in a range suggested by Hambleton and Swaminathan (1985) for a general achievement test. For the 3PLM the c parameter was allowed to range from .05 to .20 in a uniform distribution. For the 2PLM the c parameter was set to 0. The procedure was replicated 50 times.

The effect of test length and IRT model on the mean, standard deviation, skewness, and kurtosis of the indices was examined for the non-aberrant sample over the 50 replications. The means did not vary significantly; however, the skewness and kurtosis estimates were more variable over replications than for the means. In the tails of the distribution ECIZ4 was the least affected by test length and IRT model and overall the distribution of ECIZ4 most closely approximated that of the normal distribution, while the W3 was the least normal.

In order to investigate the effectiveness of the three indices, aberrant response patterns were generated for the combinations of the two test lengths, two IRT models (2PLM and 3PLM), two types of aberrance (spuriously high and spuriously low), and three levels of aberrance (10%, 15%, and 30%). Thus 24 sets of aberrant response patterns were created, each set containing 4000 examinees.

The Rudner procedure was used to generate spuriously high and spuriously low response patterns. The three appropriateness indices were then computed based on the re-estimated ability for each examinee and detection rates were calculated at four false positive rates of .01, .05, .10, and .25.

Lz was the most effective index at lower levels of aberrance and lower false positive rates. ECIZ4 was the most effective index for higher levels of aberrance and higher false positive rates. The results also showed better detection rates for the longer test, supporting the results obtained by Rudner (1983). The 2PLM was somewhat more effective than the 3PLM. As expected, increased aberrance tended to produce higher detection rates, although detection rates for 15% and 30% aberrance were quite similar.

In this study it was empirically demonstrated that factors such as test length can affect the cutoff values for the indices, even when they are standardized. This reinforces a view that for any test it may be advantageous to set cutoffs based on simulations of non-aberrant response vectors.

In this study the effectiveness of the 2PLM was compared to the 3PLM. It would also have been useful to compare other IRT models (e.g. the 1PLM). Also, in this study, the effectiveness of indices using the 2PLM was studied for response vectors that were simulated with the 2PLM, while the effectiveness of the 3PLM was studied for response vectors that were simulated with the 3PLM. However, the more complex an IRT model, the more it takes into account the characteristics of the test items. In comparisons of this type the most complex model should be used to generate response vectors, in this case the 3PLM.

Reise and Due (1991) examined the influence of test characteristics on the detection of aberrant response patterns. A model for generating aberrant response patterns was explicated based on the premise that aberrant response patterns result in less psychometric information for the individual than is predicted by the parameters of a specific model. Since test information depends on the values of the a parameter (Lord, 1980), a model in which items are differentially discriminating for different individuals was selected (Strandmark & Linn, 1987),

$$P_i(\Theta) = c + (1-c)\{ 1 + \exp[-(a_p a_i)(\Theta - b)]\}^{-1},$$

where a_i is the item discrimination, and a_p is the aberrancy level parameter for each person (p).

Systematic changes were made in the a_p parameter to generate less information than is predicted by the model. For example, when $a_p = 0.0$, the item provides no information with respect to estimating Θ.

The model was used to investigate the effectiveness of Lz for detecting aberrance as a function of the number of test items, the spread of the b parameter, and the value of the c parameter. First, the a, b, and c parameters were specified for a three parameter model. Three aberrancy levels corresponding to values for the a_p were also specified (viz. 1.0, .5, and 0.0). The $a_p = 1.0$ value produced response patterns in accordance with the null condition of no aberrance. The $a_p = .5$ condition resulted in response vectors in which the a_i parameters were reduced in half, which reduced the test information. The $a_p = 0.0$ value produced response patterns with no information. Then the model was used to generate 1,000 response vectors at each of 11 Θ values ranging from -2.5 to 2.5, in intervals of .5. The critical values for detecting aberrance were set at Lz = -1.65, which represents the .05 error rate.

Separate studies were conducted for each question. The effects of test length on detection rates were studied for tests of 7, 21, 35, and 49 items. The specifications of the item parameters was the same for all tests. The item difficulty range effects were studied for the 49 item test for five ranges: -3.0 to 3.0, -2.5 to 2.5, -2.0 to 2.0, -1.5 to 1.5, and -1.0 to 1.0. The effects of the guessing parameter were also studied on the 49 item test by specifying the c parameters to be equal to 0.0, 0.5, .10, .15, and .20 across 5 tests.

The researchers found that detection rates increased with test length and higher aberrancy levels. Detection rates were lowest when the b parameters were most clustered around the examinee's ability level. They also found that as the c parameter increased in value, detection rates decreased.

The methodology used in this study was particulary effective in explicating the effects on the detection of aberrance caused by the spread in the values of the b parameter in a test and by the increase in value of the c

parameter. The results lack some generalizability because in calculating Lz, the θ values were not reestimated and the a_i was held constant across all test items. In addition, the methodology used for generating aberrance does not provide a direct link with actual response patterns that have high or low aberrance.

Smith (1985) used a methodology which differed considerably from that which was used in the studies where only random aberrance was generated. Three reasons for aberrance were simulated on a 50 item test. For each type of aberrance 100 cases were generated based on a θ of -1 and the U1 and UB indices were computed. Based on a previous analysis, a cutoff value of 2 was set for both indices for the .05 false positive rate.

The first type of aberrance simulated was guessing (the 1PLM is based on the assumption of no guessing). For this simulation, 42 items were generated to be non-aberrant and eight items were randomly assigned a correct answer. The detection rate was 37% for U1 and 57% for UB.

The second type of aberrance that was simulated was startup disturbance, which refers to poor performance on the beginning items of a test because of test anxiety. This condition was simulated by using an ability of -2.0 to generate the responses to the first 8 items on the test and an ability of 1.0 to generate the remaining 42 items. The detection rate for U1 was 62% and 57% for UB.

The third reason for aberrance was the condition where an individual is unfamiliar with one of the content areas of a test. The test was considered to contain four content areas: items 1-12; items 13-25; items 26-37 and items 38-50. Two different abilities were used to generate the responses. The ability used to generate the responses to the first three subsets was 1.0, while the ability used to generate the responses for the fourth subset was -1.0. The analysis resulted in a 1% detection rate for U1 and a 52 % detection rate for UB.

This study was of limited scope. Important factors such as the number of items affected, the examinees' ability levels and different false positive

rates were not considered. Only two indices were compared and only the 1PLM was used. In the calculation of the UB index the items in the test are divided into sets of items. The investigator expected differences in performance on these sets. The results of this study suggest that the UB index may be superior to the U1 index when different sets of items are generated with different difficulty levels and the UB index corresponds exactly to these sets. However, without a more comprehensive design in which a greater number of factors are taken into account, no clear conclusions can be reached from this study.

Gafni (1987) investigated the performance of C^*_1, ECIZ2, ECIZ4, W3, U3, Lz, and UB3 in detecting both random and systematic forms of aberrance, using the 3PLM. The simulation involved a 200 item test with parameters based on the Armed Services Vocational Aptitude Battery (ASVAB) Arithmetic Reasoning (AR) items. Each of the indices was evaluated under conditions which varied the form of aberrance, the percent of total number of items subjected to aberrance (10%, 30%, and 50%), and the level of ability.

Five forms of aberrance were modeled based on different causes for aberrance (2,500 response vectors were generated for each form). In Form 1, all of the test items were subjected to the Levine technique for generating spuriously low response vectors. In Form 2, the Levine technique for generating low aberrance was applied for a specified percentage of "unreached" items at the end of the test. This form of aberrance might be produced by a plodder (someone who takes too much time at the beginning of a test). In Form 3, the items in the test were divided into sets and an examinee was given a different ability level for different sets of items on the test. This form of aberrance might be produced on a test which is divided into subtests or subgroupings of items, where an examinee is unfamiliar with certain concepts underlying one or more of the subgroups. In Form 4, the Levine technique for generating low aberrance was applied to "unreached" items on a test which had items ordered from easiest to most difficult. This was exactly the same as Form 2, but in Form 2 the test items were not ordered according to difficulty.

In Form 5, the Levine technique for generating low aberrance was applied to different sets of items on the test. This form of aberrance is very similar to Form 3 aberrance. The causes cited for Form 3 and Form 5 aberrance are exactly the same.

In order to examine the relationship between the indices and examinee ability, the response vectors of 5,000 non-aberrant examinees were simulated. Two hundred examinees were generated at each of 25 ability levels, in the range -3 to 3, in intervals of .25. Correlations of the index values and ability were calculated. In addition, in order to detect a possible non-linear trend, the means of the indices computed for each of the 25 θ intervals were plotted against θ. The C^*_1 was the most correlated with ability (r= -.77), while the U3 was less highly correlated (approximately .17). The rest of the indices showed no linear relationship to ability. None of the indices showed a clear non-linear trend except for U3.

The effectiveness of the indices, examined using cutoff values obtained from the non-aberrant sample at the .01, .05, and .10 false positive rates, seemed to depend on the form of the aberrance and the number of items subjected to aberrance (C^*_1 was used only for the Form 1 aberrance, while the rest of the indices were used for all of the forms). Lz was the most effective for the detection of forms 1, 2, and 5 aberrance. For Form 3 aberrance (different ability underlying responses to different parts of the test) and Form 4 (random responses for unreached items when the items were ordered according to difficulty), the appropriate UB3 index (i.e. the one in which the sets used for calculating the index corresponded to the sets used for the items) was the most effective.

Detection rates were quite high. Most forms of aberrance were detected perfectly with no error when examinees at the high levels of ability were subjected to aberrance; aberrance for examinees at the low levels of ability was generally not well detected.

This is the only research on the effectiveness of a variety of appropriateness indices for detecting systematic aberrance. While the above

findings are encouraging, they are based on only one 200 item test. Tests of this length are usually broken down into shorter subtests and appropriateness indices are applied to the subtests. It is questionable whether the results are generalizable to other shorter tests.

The finding that UB3 was more effective than U3 for aberrance involving different ability underlying responses to different parts of the test corresponds to the results obtained by Smith (1985). In this form of aberrance it is realistic that the item sets used for calculating the UB indices correspond exactly to the sets used for differences in the ability in the simulation, as this can be predicted by item formats and subtests. However, this correspondence is not appropriate for Form 2 (random responses for unreached items) and Form 4 aberrance (random responses for unreached items when the items are ordered according to difficulty) because there is very little way of knowing the point in the test at which examinees will run out of time on the test. Thus the effectiveness of the UB3 index for detecting these forms of aberrance under real conditions is questionable.

Summary and Future Research

The research on appropriateness indices has answered a number of important questions, while others remain to be explored more fully. This section summarizes this research and presents issues for future research.

Distribution of the indices. Most indices, in their original form, demonstrate a linear or non-linear relationship to the total score on the test and, when possible, they have been standardized. In order to standardize an index it is necessary to obtain estimates of its conditional mean and variance. To date none of the heuristic indices have been standardized. Consequently, all of the heuristic indices are quite strongly related to the total score on the test.

The standardization of the IRT indices has been successful to a point. In particular, the ECIZ2, ECIZ4 (Tatsuoka & Tatsuoka; 1982, Drasgow, Levine, & McLaughlin, 1987; Gafni 1987), and the Lz (Drasgow, Levine, & McLaughlin,

1987; Gafni, 1987; Noonan, 1989) indices have been found to have relatively little relationship to the ability level of the examinee.

The standardization of the indices may be of limited practical value. Smith (1982) has noted that each combination of test length and distribution of item difficulties and person abilities represents an untested situation. Molenaar and Hoijtink (1990) have argued that the distributions of the indices under the null hypothesis of no aberrance is different for different latent trait values and have recommended that for longer tests, these distributions be determined with a Monte Carlo approach.

Noonan (1989), in the most comprehensive study to date on the distributions of the Lz, ECIZ4, and W3, showed that while the mean values of the Lz, ECIZ4, and W3 were not greatly affected by test length and IRT model, the skewness and kurtosis of the indices varied considerably. In practical applications it is the cutoff values of the indices that are most important.

While it may be best to simulate the test in question under conditions of non-aberrance to obtain cutoff values at various false positive rates, the advantages of using this technique for obtaining cutoff values must be weighed against practical considerations. Obtaining these cutoff values is both complex and time consuming. For two of the standardized indices, Lz and ECIZ4, it may be possible to use values based on the normal (0,1) distribution. Drasgow, Levine, and McLaughlin (1987) and Noonan (1989) found that both of these indices followed the normal (0,1) distribution quite closely, while Drasgow, Levine, and Williams (1985) found that the Lz followed the normal (0,1) distribution quite closely. It is possible that this alternative approach may produce similar results, but no researcher has compared the two approaches.

Effects of increased test length on detection rates The research that has been conducted to date suggests that when an equal percentage of the items in two tests are aberrant, the detection rates for the longer test are better than those of the shorter test. Rudner (1983) found that the detection rates on an 80 item test were higher than those on a 45 item test. Drasgow, Levine,

and McLaughlin (1987) also had better detection rates on a longer (85 item) test than on a shorter (30 item) test. However, in both of these studies the two tests were different and this may have confounded the results.

Noonan (1989) compared the detection rates for a 40 item test and an 80 item test, which had both been generated using similar item parameters. The detection rates for the 80 item test ranged from 21% to 71% across the different levels of aberrance and false positive rates, while the corresponding range for the 40 item test was from 12% to 62%.

Reise and Due (1990) compared the detection rates of four tests ranging in length from 7 to 49 items. Detection rates increased with test length.

With the advent of computer adaptive testing, it may be important to have an understanding of the effectiveness of the indices for tests that are very short (perhaps 7-15 items). However, all else being equal, one would expect a longer test to provide more information than a shorter test and thus better detection rates. The research findings seem to support this.

Effect of the amount of aberrance on detection rates Results of studies showing the relationship of the degree of aberrance to the rate of detection suggest that as the amount of aberrance increases detection increases, but the increase is less at higher rates of aberrance. Drasgow, Levine, and Williams (1985) induced 10%, 20%, and 30% aberrance on an 85 item test and found a larger increase in the detection rate between 10% and 20% aberrance than between 20% and 30% aberrance. Levine and Rubin (1979) simulated 4%, 10%, 20%, and 40% aberrance and found the greatest increase in detection occurred when the increase was from 4% to 10%. Drasgow and Levine (1986) simulated aberrance at the 15% and 30% level and found the detection rate higher at the 30% level. Rudner (1983) actually changed the responses to 6%, 13%, 19%, and 25% of the items on an 80 item test and 11%, 28%, and 33% on a 45 item test. For the shorter test, the increase in detection rate was about equal from one level of aberrance to the next. For the longer test the increase was greatest from 15% to 20%. Noonan (1989) changed 10%, 15%, and 30% of the items on a 40 item test and on an 80 item test. The increase in

detection was greater from 10% to 15% than from 15% to 30%. Gafni (1987) found that as the proportion of items subjected to aberrance was increased from 10% to 30% to 50% the detection rate increased, but the increase was larger between the first two levels than between the last two.

It seems plausible to expect that as the amount of aberrance in a test increases the detection rate of the appropriateness indices would increase and the research has consistently borne this out. It is important to note, however, that a methodological problem exists in the detection of simulated aberrance. As the percentage of the simulee's responses in a test that are made aberrant increases, the original ability level is changed more and more. At some point it no longer resembles the original ability level. As this point is approached it is difficult for the detection rate to increase because the estimate of ability level which is used for calculating the indices has become increasingly affected by the aberrance.

Effectiveness of the indices There is considerable evidence that Lz is one of the most effective of the appropriateness indices (Drasgow, Levine, & Williams, 1985; Drasgow, Levine, & McLaughlin, 1987; Gafni, 1987; Noonan, 1989).

Of the standardized extended caution indices, ECIZ2 and ECIZ4 have been the most researched and appear to be similar in effectiveness. In comparisons with other indices both of these indices have proven to be amongst the most effective (Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, McLaughlin, & Candell, 1987; Noonan, 1989).

The UB index was found to be the most effective index for certain forms of systematic aberrance, but requires one to assume which items may be aberrant (Smith, 1985; Gafni, 1987). This index should be included in future studies of systematic aberrance when these assumptions are considered to have been met.

The zh and IOV indices are the only practical polychotomous indices that have been developed to date; they have not been very effective. Other polychotomous indices should be developed and examined in future research.

The heuristic indices have not been very effective in detecting aberrance in comparison to the IRT indices, but this may be because they are not standardized. Heuristic indices could be used when the assumptions of the model based indices are not met and it might be worth investigating whether or not they can be as effective as the IRT based indices and under what circumstances. In research on the heuristic indices, C^*_1 is the index of choice because it is the most well known of the heuristic indices and because it is conceptually related to the extended caution indices.

The U3 and W3 indices have not been consistently effective in detecting aberrance (Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, McLaughlin, & Candell, 1987; Gafni, 1987; Noonan, 1989).

In sum, the Lz, ECIZ2, and ECIZ4 indices have proven to be the most effective of the appropriateness indices in a variety of circumstances and are among the most practical to use. These indices should be considered in future research studies.

Detection of systematic aberrance Little research has been conducted on the effectiveness of the indices for detecting systematic aberrance. Smith (1985) examined the effectiveness of the U1 and UB using the 1PLM. However, only one θ level was used. Gafni (1987), using the 3PLM, examined the effectiveness of the C^*_1, ECIZ2, ECIZ4, W3, U3, Lz, and UB3 indices. However, the study only involved one long test of 200 items. The sets of items used in the calculation of the UB3 index always corresponded exactly to the sets of items used for the simulated aberrance, which is not realistic for some forms of aberrance.

Future research is necessary to determine the effectiveness of a variety of appropriateness indices for systematic aberrance and with shorter tests. In addition, for the forms of aberrance where it is unrealistic to predict the sets of items that will be affected, the UB should not be used.

The effectiveness of different IRT models Drasgow (1982) and Noonan (1989) compared the effectiveness of using different IRT models for

calculating appropriateness indices. More research is needed to establish if there are distinct advantages to using any one model.

The effect of sample size The effect of using different sample sizes for the estimation of item parameters has been researched only once for the 1PL and 3PL models (Drasgow, 1982). More research is necessary to determine the effect of sample size for the 1PL, 2PL, and 3PL models.

The effects of test characteristics on detection rates The effects of test characteristics on detection rates has only been studied once using a methodology that modelled the aberrance indirectly (Reise & Due, 1990). In that study an increase in the variance of the b parameters for the test items and a decrease in the values of the c parameters were both found to increase detection rates. In future research, the effects of these conditions should be examined using the more direct simulation approach of actually modifying non-aberrant response vectors. In addition, it would be useful to examine the effects that the level of the a parameter has on detection rates.

Research Problem

Research has shown that Lz and ECIZ4 are two of the most useful appropriateness indices for a number of reasons. First, they are considered useful because they have had higher detection rates than other indices. Second, these indices have been shown to have little relationship to ability, which allows for a clear interpretation of their values. Third, their distributions have shown some evidence of approximately following a standard normal distribution. This enables test practitioners, should they wish, to use a standard normal distribution for setting the cutoff values of these indices at various false positive rates. However, a number of questions remain regarding the practical use of these indices. The purpose of this research is to examine three of these questions in relation to the Lz and ECIZ4 indices. The first problem concerns the effect of different test conditions on the cutoff values of the indices. It has been noted that the distribution of these indices may be affected by different test conditions,

but little research has been conducted to examine this issue. For practical purposes, it is important to understand how the distributions of the indices are affected in the extremes of their distributions, where cutoff values for various false positive rates are selected. No extensive research has been conducted to determine if certain test conditions have systematic effects on these values. In this research the effects on the cutoff values for different distributions of the b item parameter in a test, the level of the a item parameters and the sample size used to estimate item parameters are examined. The effects of these three variables have never been examined directly. In addition the effects of using the 2PL or 3PL models is researched. Only one researcher has examined this problem (Noonan, 1989). However, the methodology did not involve the use of estimated item parameters which are, of course, necessary in practical situations. Therefore, the first question posed is Are the cutoff values of Lz and ECIZ4 affected by the range of the b parameters, the level of the a parameters, the IRT model, and the sample size used to estimate item parameters?

The second problem addressed in this research concerns the conditions under which test practitioners are likely to obtain the highest detection rates. It is important for the test practitioner to understand how certain variables affect detection rates. The variables that are of concern are the same as those that are examined for their effects on cutoff values. The first variable is the range of the b parameter in the test. Although Reise and Due (1991) have shown that the variance of the b parameters in a test has an effect on the detection rates for aberrance, in their study the aberrance was modelled indirectly. To determine the magnitude of the effect and whether the effect is the same for high and low aberrance, it is necessary to directly simulate the aberrance. The second variable, the level of the a parameters in a test, has never been researched. The third variable, the IRT model, has not been fully researched for the 2PL and 3PL models. Noonan (1989) compared the effectiveness of the two indices using the 2PL and 3PL models. However, the data were generated to fit both models and the item parameters used for

calculating the indices were not estimated. This is not very realistic for practical test situations. Guessing is likely to occur even when the 2PLM is used, and the item parameters must be estimated. Finally, the fourth variable, the effect of sample size used to estimate item parameters, has only been researched once and never with the 2PLM. To examine the second problem described above, the following specific question is posed: How effective are the Lz and ECIZ4 indices at detecting low (viz. spuriously low) and high (viz. spuriously high) aberrance under combinations of range of the b parameter, level of the a parameter, IRT model, and sample size?

The third problem concerns treating the indices as if they are distributed as a standard normal distribution for purposes of selecting cutoff values for the indices at various false positive rates. Although some research has shown that both indices approximately follow a standard normal distribution, the practical effects of using cutoff values from a standard normal distribution instead of setting the cutoff values based on the simulation of non-aberrant response vectors has never been examined. This question is of great practical importance to test practitioners. If a standard normal distribution can be used for setting cutoff values, it reduces the time and effort of using the indices and thus makes them more practical to use. To examine this problem the following specific question is posed: What are the consequences of using cutoff values obtained from the standard normal distribution as compared to using cutoff values obtained by simulating a nonaberrant sample of response vectors for the test in question?

In the next chapter the methodology used to investigate these three problems is presented and discussed.

CHAPTER IV

METHODOLOGY

The purpose of this research was to examine the following questions: (1) Are cutoff values for Lz and ECIZ4 affected by the range of the distribution of the b parameter, the level of the a parameter, IRT model, and sample size? (2) How effective are the Lz and ECIZ4 indices at detecting low and high aberrance under combinations of degree of variance of the b parameter, level of the a parameter, IRT model, and sample size used to estimate item parameters? (3) To what extent are the detection rates obtained by using cutoff values obtained from the standard normal distribution similar to those obtained by using cutoff values obtained by simulating a nonaberrant sample of response vectors for the test in question?

A Monte Carlo simulation technique was used to generate the data for this study because it is the only practical way to represent all of the conditions required for this research in all of their combinations. In addition, the Monte Carlo technique allows the conditions to be replicated and enables the researcher to know exactly the true amount of aberrance that exists in the response vectors. Also, with replications it is possible to examine the stability of results. None of this is possible with actual test data.

This chapter is divided into two sections. In the first section the independent variables are presented and discussed. In the second section the procedures for generating and analyzing the data are presented and discussed.

The Independent Variables

The first independent variable generated for this study was the variance of the b parameter across the test items, which was labelled "Diff". Two levels were used for this variable: "high Diff", represented by test items with b values ranging from -2.5 to 2.5, uniformly distributed, and "low Diff" with b values ranging from -1.5 to 1.5, uniformly distributed. There was some expectation that the high condition might result in higher detection rates than for the low condition. For high Diff there is information for a wider range of examinee abilities and the non-aberrant distribution of the indices is better estimated. Also, for the high condition the examinees have the chance of responding inappropriately to items that are farther from their level of ability. These inappropriate responses are very unlikely and so provide a clearer indication that the examinee has responded aberrantly. Furthermore, in the one study where this question has been examined, using a different methodology, the high Diff had resulted in better detection (Reise & Due, 1990).

The second independent variable was the level of the a item parameters within the test, which was labelled "Disc". Two levels were used for this variable: "high Disc", representing a parameters ranging from .8 to 1.4, in a uniform distribution and "low Disc", representing a parameters ranging from .5 to 1.1, in a uniform distribution. There was no expectation as to which condition would affect cutoff values or result in better detection rates.

The third independent variable was the IRT model. Two models were used for this study, the 2PLM and the 3PLM. There was no expectation as to which condition would result in better detection rates. The effectiveness of IRT models for appropriateness indices have only been compared in two studies. In one study the 1PLM and the 3PLM were compared (Drasgow, 1982). The 3PLM

was more effective. In the second study the 2PL and 3PL models were compared (Noonan, 1989). The 2PLM was more effective, but the 2PLM had been fitted to data generated with the 2PLM. One might expect the 3PLM to perform better because, as seen in the next section, in this study the data were generated using the 3PLM. Thus better estimates of examinee ability should be obtained. On the other hand, guessing introduces noise and aberrance can get confused with guessing. Also, the c item parameter is difficult to estimate accurately and to the extent that the c item parameter is poorly estimated, the 3PLM may perform more poorly.

The fourth independent variable was the sample size of the responses vectors, labelled "Sampsiz". The LOGIST program was used to estimate items parameters in this study. Using this program for a 60[3] item test, it has been suggested that sample sizes between 1000 and 2500 can reasonably be employed for both the 2PL and 3PL models (Hulin, Lissak, & Drasgow, 1982). However, the estimates of the item parameters are less accurate with smaller sample sizes. There was no expectation regarding the effects of these two sample sizes on detection rates or cutoff values. Only one comparison of the effect of sample size on the effectiveness of appropriateness indices had been conducted (Drasgow, 1982). In that study, for low aberrance, the 3PLM performed better that the 1PLM even when a sample size of 500 was used.

Procedures for Data Generation and Analysis

Because the distribution of appropriateness indices may be affected by a number of factors, several researchers have recommended that practitioners actually simulate a non-aberrant sample of examinees to obtain the most

[3]In this study a 60 item test was simulated.

accurate estimate of cutoff values at various false positive rates. In this study, the detection rates obtained using these cutoff values were used in determining the effectiveness of the indices under the various conditions. For purposes of comparison, detection rates were also calculated using cutoff values obtained from the standard normal distribution.

In this study a 60 item test was simulated under various conditions. A test length of 60 items was chosen because it is a common length for standardized, multiple-choice, paper and pencil tests. These tests can usually be administered in a reasonable amount of time and a reasonable degree of test reliability can be expected. In addition, past research on detection rates shows that appropriateness indices can achieve reasonable detection rates with tests of 60 items. Furthermore, 60 item tests tend to be about the median test length used in most of the studies on appropriateness indices. The procedure for the study involved a series of nine steps.

Step 1: Data generation Binary response data were generated according to the 3PLM. The 3PLM was chosen because it is typically found that a c parameter is appropriate for multiple-choice tests. The data were generated for eight conditions, representing two levels for Diff, Disc, and Sampsiz. Examinee ability values were generated from the normal (0,1) distribution. The c parameter was generated to range from .05 to .20 in a uniform distribution. The item parameters, ability levels, and response vectors were generated using the computer program, DATAGEN (Hambleton & Rovinelli, 1973) which was updated by Carlson (1985).

Step 2: Parameter estimation The item parameters were estimated for each of the eight conditions created in the first step using the LOGIST program (Wingersky, 1983; Wingersky, Barton, & Lord, 1982). To introduce the effect

of model, the estimates were made twice, first using the 2PLM and second using the 3PLM.[4]

Step 3: Cutoff values The Lz and ECIZ4 indices were calculated using 2000 response vectors generated according to the combinations of Diff and Disc conditions in Step 1. The two indices were calculated for the 2PL and 3PL models using the item parameter estimates from Step 2. A program designed by Drasgow (1987) was used to calculate the values of the indices.

Since, false positive rates above .10 are not likely to be useful in real situations, the values of the indices at the .01, .05, and .10 false positive rates were obtained (i.e. at the 99th, 95th, and 90ieth percentiles for the ECIZ4 and at the 1st, 5th, and 10th percentiles for the Lz).

Step 4: Generation of aberrant response vectors In this step, 1500 spuriously high and 1500 spuriously low response vectors were generated for each of the 1€ combinations of conditions, using the parameter specifications from step 1. To generate the spuriously high condition, non-aberrant response vectors were generated with ability levels ranging from -3.00 to 0 in a uniform distribution; for the spuriously low condition response vectors were generated with ability ranging from 0 to 3.00 in a uniform distribution. The ranges for the ability were chosen because past studies have shown that detection rates are better for the extremes and aberrant response patterns are not readily identified in the remaining half of the distribution.

For the spuriously high condition, for each vector 10 incorrect responses were randomly selected and changed to be correct. For the spuriously low condition, for each vector 10 correct responses were randomly

---

[4]It is noted that the samples used for the Model variable were not independent. This increases the chance of finding differences for this variable.

selected and changed to be incorrect (i.e. for each of the conditions, 16.6% of the items were changed).

Step 5: Detection rates For each of the 16 combinations, for high and low aberrance Lz and ECIZ4 were calculated using estimated item parameters from Step 2 (for the calculation, ability estimates were reestimated in the modified response vectors).

For each of the two indices, detection rates were calculated for each of the 32 conditions and for each of the three false positive rates using the cutoff values obtained in step 3, and also using cutoff values from the normal (0,1) distribution.

Step 6: Replication To avoid chance error, the process described in the above five steps was repeated 90 times. For each replication, new seed numbers were used to generate the a and b item parameters and for generating the abilities.

Step 7: Effects of the independent variables on cutoff values For each of the false positive rates, the means and standard deviations of the cutoff values of the two indices were calculated over the 90 replications. The means of the cutoff values were used to compare the cutoff values of the two indices to each other and to the cutoff values from the standard normal distribution. The standard deviations were used to compare the stability of the cutoff values of the two indices over replications.

ANOVAS were conducted to examine the effects of the four independent variables at each cutoff, for each index. This resulted in three ANOVAS for each of the two indices. Typically significance tests are conducted at the .05 level. Since there were 6 ANOVAS overall, the Bonferroni procedure was used to adjust the significance level so as to avoid capitalizing on chance. The basic idea of this procedure is to divide the alpha level selected by

the number of tests performed (Rosenthal and Rosnow, 1991). Thus, for each of the above ANOVAS the adjusted significance level was .008 (i.e. .05/6).

Step 8: Effects of independent variables on detection rates The mean detection rate and standard deviation for each of the combinations of independent variables was calculated over the 90 replications. To examine the effects of the independent variables on the detection rates, a series of four way between-groups ANOVAS were conducted, using the test model (2 or 3 PLM), the condition of the a parameter (high or low), the condition of the variance of the b parameter (high or low) and the sample size used to estimate the item parameters (1000 or 2500 examinees) as the independent variables, with the detection rate for the index as the dependent variable.[5]

Initially, it was reasoned that if the ANOVAS were performed using the detection rate for aberrant response vectors at each of the three false positive rates, the ANOVAS would not be independent. This is because the detected response vectors at the .01 false positive rate would be counted again for the .05 rate and for the .10 rate. Likewise the response vectors for the .05 rate would be counted again for the .10 rate. To avoid this situation of overlapping variance, the proportions of the detected response vectors were calculated for three intervals. The first interval, corresponded directly to the .01 false positive rate. The second interval represented the proportion of those detected response vectors falling between the .01 and .05 false positive rates. The third interval represented the proportion of those detected response vectors falling between the .05 and .10 false positive rates. As a preliminary analysis a

[5] The use of a data analysis strategy in which a MANOVA precedes multiple ANOVAS was considered inappropriate for the research question posed in this study (Huberty & Morris, 1989).

series of 12 ANOVAS were performed using the detection rates of the indices in the three intervals.

However, this analysis was not reported for two reasons. The first reason relates to analysis of the results based on the intervals. In the first interval, there was a tendency for the effects of the independent variables to be consistent for both indices and for high and low aberrance. However, the effects for the second and third intervals were inconsistent, complex, and confusing. For example, for high aberrance a three way interaction of Diff by Disc by Model occurred for ECIZ4 in the second and third intervals. However, the interaction was different for each of the intervals. As a second example, for low aberrance in the second interval a disordinal Diff by Model interaction occurred for Lz and a disordinal Disc by Model interaction occurred for ECIZ4. Neither of these two interactions occurred for the other two intervals. Even when no interactions took place or where the interactions were ordinal, the results were confusing because the effects of the independent variables in the second and third intervals tended to be the opposite of the effects for the first interval (This may have resulted because the major portion of the aberrant response vectors that were detected were in the first interval.). The second problem with the analysis based on the interval is that it is not directly related to cutoff values. Test practitioners are most interested in the direct effects of the variables on detection rates at the .01, .05, and .10, false positive rates.

For these reasons, for each of the two indices six ANOVAS were conducted (one for each of the three false positive rates for high aberrance and one for each of the three false positive rates for low aberrance). The significance level for each ANOVA was set at .004 after the adjustment according to the Bonferroni procedure (i.e. .05/12).

Step 9: Detection rates using cutoff values from the N(0,1) distribution

The detection rates obtained using cutoff values from the standard normal distribution were compared to those obtained using cutoff values from the simulated response vectors.

In the next chapter the results of this study are presented and discussed.

CHAPTER V

RESULTS AND DISCUSSION

This chapter is divided into three sections. In the first section, the cutoff values for false positive rates of .01, .05, and .10 are examined. Their stability over replications, their correspondence to tabled values of a standard normal distribution, and the effects of the independent variables are presented and discussed. In the second section the effects of the independent variables on the detection rates are presented and discussed. In addition, detection rates for high and low aberrance are compared, as well as detection rates for the two indices. In the third section the effects of assuming a standard normal distribution for setting cutoff values is investigated. Finally, a summary of the results and discussion is presented.

Cutoff Values

Three issues are addressed in this section. The first concerns the stability of the cutoff values. The second issue is the correspondence of the observed cutoff values to those of a standard normal distribution. The third issue is the effect of the independent variables on cutoff values.

Stability

For each of 90 replications, 2000 nonaberrant response vectors were generated for each of 16 combinations of the independent variables and values of Lz and ECIZ4 were calculated. The mean and standard deviation of Lz at the .01, .05, and .10 false positive rates over the 90 replications are shown in Table 4. Similarly, the mean and standard deviation of ECIZ4 are presented in Table 5.

Tables 4 and 5 were examined to determine the stability of the cutoff values - first over the different combinations of conditions of the independent variables and second over replications. The mean of the cutoff

Table 4

Mean and Standard Deviation of Lz Values at .01, .05, and .10 False Positive Rates Over 90 Replications

CONDITION				FALSE POSITIVE RATE					
				.01		.05		.10	
Sampsiz	Variance of Diff.	Disc.	Model	M	SD	M	SD	M	SD
1000	low	low	2PLM	-2.134	.114	-1.414	.074	-1.061	.061
1000	low	low	3PLM	-2.282	.118	-1.528	.070	-1.155	.061
1000	low	high	2PLM	-2.253	.128	-1.506	.084	-1.132	.073
1000	low	high	3PLM	-2.380	.120	-1.595	.077	-1.198	.065
1000	high	low	2PLM	-2.519	.139	-1.686	.095	-1.272	.080
1000	high	low	3PLM	-2.504	.133	-1.697	.086	-1.290	.065
1000	high	high	2PLM	-2.752	.169	-1.809	.097	-1.336	.082
1000	high	high	3PLM	-2.550	.127	-1.722	.073	-1.302	.054
2500	low	low	2PLM	-2.115	.107	-1.412	.077	-1.058	.065
2500	low	low	3PLM	-2.237	.110	-1.498	.076	-1.124	.062
2500	low	high	2PLM	-2.219	.138	-1.470	.072	-1.103	.059
2500	low	high	3PLM	-2.340	.127	-1.572	.072	-1.180	.059
2500	high	low	2PLM	-2.493	.127	-1.668	.075	-1.254	.060
2500	high	low	3PLM	-2.479	.120	-1.682	.065	-1.273	.060
2500	high	high	2PLM	-2.738	.166	-1.796	.094	-1.324	.076
2500	high	high	3PLM	-2.553	.123	-1.710	.071	-1.291	.056
Entire Sample				-2.409	.230	-1.610	.147	-1.210	.112
MARGINALS									
Sampsiz	1000			-2.42	.23	-1.62	.15	-1.22	.11
	2500			-2.40	.23	-1.60	.15	-1.20	.11
Variance of Diff.	low			-2.24	.15	-1.50	.10	-1.13	.08
	high			-2.57	.17	-1.72	.10	-1.29	.07
Disc.	low			-2.35	.20	-1.57	.14	-1.19	.11
	high			-2.47	.24	-1.65	.15	-1.23	.11
Model	2PLM			-2.40	.28	-1.60	.18	-1.19	.13
	3PLM			-2.42	.17	-1.63	.11	-1.23	.09

Table 5

Mean and Standard Deviation of ECIZ4 Values at .01, .05, and .10 False Positive Rates Over 90 Replications

CONDITION				FALSE POSITIVE RATE					
				.01		.05		.10	
Sampsiz	Variance of Diff.	Disc.	Model	M	SD	M	SD	M	SD
1000	low	low	2PLM	2.451	.086	1.730	.057	1.346	.051
1000	low	low	3PLM	2.403	.102	1.684	.058	1.308	.052
1000	low	high	2PLM	2.435	.104	1.673	.056	1.298	.047
1000	low	high	3PLM	2.354	.101	1.628	.052	1.266	.052
1000	high	low	2PLM	2.510	.107	1.736	.075	1.334	.063
1000	high	low	3PLM	2.436	.102	1.671	.062	1.292	.052
1000	high	high	2PLM	2.629	.122	1.766	.070	1.326	.058
1000	high	high	3PLM	2.400	.104	1.639	.070	1.255	.056
2500	low	low	2PLM	2.395	.077	1.676	.053	1.303	.047
2500	low	low	3PLM	2.346	.102	1.626	.056	1.259	.049
2500	low	high	2PLM	2.378	.105	1.640	.050	1.265	.041
2500	low	high	3PLM	2.291	.106	1.573	.060	1.206	.051
2500	high	low	2PLM	2.474	.092	1.702	.061	1.299	.043
2500	high	low	3PLM	2.388	.106	1.638	.055	1.250	.047
2500	high	high	2PLM	2.560	.109	1.732	.068	1.304	.051
2500	high	high	3PLM	2.387	.100	1.615	.062	1.228	.045
Entire Sample				2.427	.130	1.670	.079	1.284	.063
MARGINALS									
Sampsiz	1000			2.45	.13	1.69	.08	1.30	.06
	2500			2.40	.13	1.65	.08	1.26	.06
Variance of Diff.	low			2.38	.11	1.65	.07	1.28	.06
	high			2.47	.13	1.69	.08	1.29	.06
Disc.	low			2.43	.11	1.68	.07	1.30	.06
	high			2.43	.15	1.66	.09	1.27	.06
Model	2PLM			2.48	.13	1.71	.07	1.31	.06
	3PLM			2.38	.11	1.63	.07	1.26	.06

values of the indices varied considerably over the different combinations of the independent variables for each of the false positive rates. For example, at the .01 false positive rate, the mean cutoff value for Lz ranged from -2.115 to -2.752, while the mean cutoff value for ECIZ4 ranged from 2.291 to 2.629. A comparison of the ranges of the mean cutoff values for the two indices shows that over the combinations of the independent variables, the ECIZ4 had a range that was approximately one half of the range of Lz. Lz had a range of .637, .397, and .278 at the .01, .05, and .10 false positive rates, while the corresponding values for ECIZ4 were .338, .193, and .140.

The marginal values of the cutoff values in Tables 4 and 5 were examined for the magnitude of the effects of the independent variables. For Lz, Diff, and Disc had the greatest effect. High Diff produced more extreme values than low Diff. The cutoff values for high Diff were .33, .22, and .16 more extreme than for low Diff at the .01, .05, and .10 false positive rates, respectively. The cutoff values for high Disc were .12, .08, and .04 more extreme than for low Diff at the .01, .05, and .10 false positive rates. For ECIZ4, the Diff and Model variables had the greatest effects, but the magnitude of the effects were less than was seen for Lz. High Diff produced more extreme values than low Diff. The cutoff values for high Diff were .09, .04, and .01 more extreme than for low Diff at the .01, .05, and .10 false positive rates, respectively. The 2PLM produced more extreme values than the 3PLM - .10, .08, and .05 at the .01, .05, and .10 false positive rates, respectively.

The stability of the estimates of the cutoff values over replications was determined by examining the standard deviation of the cutoff values over replications. As expected, for both indices the standard deviations were greater at the lower false positive rates. In addition, it was found that for all three false positive rates the standard deviations for ECIZ4 were approximately one half of those for Lz. Thus, ECIZ4 was more stable over replications as it was over experimental conditions.

Comparison of cutoff values

In order to compare the cutoff values for the two indices to those obtained from the standard normal distribution, the difference between the mean values obtained from the generated response vectors and the values of the standard normal distribution are presented in Table 6 . Under the standard normal distribution, the values for Lz at the .01, .05, and .10 false positive rates are -2.325, -1.645, and -1.282, respectively. The corresponding values for ECIZ4 are 2.325, 1.645, and 1.282. In Table 6, observed cutoff values that are more extreme than those of the standard normal distribution are shown in brackets, while those that are less extreme have no brackets.

The range in the extremeness of the cutoff values was quite high for both indices. Of course, the range was the same as for the range of the cutoff values over the combinations of conditions, which is presented above. Over the three false positive rates, there was a very slight tendency for Lz cutoff values to be less extreme than those of the standard normal distribution (however, Lz tended to be more extreme at the .01 false positive rate). Over the 16 combinations of conditions and three false positive rates, the cutoff value for Lz was less extreme in 52% of the cases. Over the entire sample, the cutoff value for Lz was more extreme than that of the standard normal distribution at the .01 false positive rate, but less extreme at the .05, and .10 false positive rates. Table 6 shows a clear tendency for the cutoff values of ECIZ4 to be more extreme than those of the standard normal distribution. Over the 16 combinations of conditions and three false positive rates, the cutoff value for ECIZ4 was more extreme than that of the standard normal distribution in 66% of the cases. Over the entire sample, the cutoff values for ECIZ4 were more extreme that those of the standard normal at all three false positive rates.

The proximity of the cutoff values to those of a standard normal distribution were compared for the two indices. Over the 16 combinations of conditions, at the three false positive rates, the cutoff values for ECIZ4

Table 6

Difference Between Mean of Observed Cutoff Values and Values Under the Standard Normal Distribution at Three False Positive Rates

CONDITION				Lz			ECIZ4		
Sampsiz	Diff	Disc	Model	.01	.05	.10	.01	.05	.10
1000	low	low	2PLM	.191	0.231	0.221	(.126)	(.085)	(.064)
1000	low	low	3PLM	.043	0.117	0.127	(.078)	(.039)	(.026)
1000	low	high	2PLM	.072	0.139	0.150	(.110)	(.028)	(.016)
1000	low	high	3PLM	(.055)	0.050	0.084	(.020)	.017	.016
1000	high	low	2PLM	(.194)	(0.041)	0.010	(.185)	(.091)	(.052)
1000	high	low	3PLM	(.179)	(0.052)	(0.008)	(.111)	(.026)	(.010)
1000	high	high	2PLM	(.427)	(0.164)	(0.054)	(.304)	(.121)	(.044)
1000	high	high	3PLM	(.225)	(0.077)	(0.020)	(.075)	.006	.027
2500	low	low	2PLM	.210	0.233	0.224	(.070)	(.031)	(.021)
2500	low	low	3PLM	.088	0.147	0.158	(.021)	.019	.023
2500	low	high	2PLM	.106	0.175	0.179	(.053)	.005	.017
2500	low	high	3PLM	(.015)	0.073	0.102	.034	.072	.076
2500	high	low	2PLM	(.168)	(0.023)	0.028	(.149)	(.057)	(.017)
2500	high	low	3PLM	(.154)	(0.037)	0.009	(.063)	.007	.032
2500	high	high	2PLM	(.413)	(0.151)	(0.042)	(.235)	(.087)	(.022)
2500	high	high	3PLM	(.228)	(0.065)	(0.009)	(.062)	.030	.054
Entire Sample				(.084)	0.035	0.072	(.102)	(.025)	(.002)
MARGINALS									
Sampsiz		1000		(.09)	.04	.06	(.12)	(.04)	(.02)
		2500		(.07)	.06	.08	(.07)	.00	.02
Diff		low		.09	.16	.15	(.05)	.00	.00
		high		(.24)	(.06)	(.01)	(.14)	(.04)	(.01)
Disc		low		(.02)	.09	.09	(.10)	(.03)	(.02)
		high		(.14)	.01	.05	(.10)	(.01)	.01
Model		2PLM		(.07)	.06	.09	(.15)	(.06)	(.03)
		3PLM		(.09)	.03	.05	(.05)	.02	.02

were closer to those of a standard normal distribution in 79% of the cases. The marginal cutoff values in Table 6 were also clearly closer for ECIZ4. At the .01, and .10 false positive rates, the values of ECIZ4 were closer in 86% of the cases, while at the .05 false positive rate ECIZ4 was closer or the same in 63% of the cases.

In order to determine if the cutoff values departed significantly from the tabled values of a standard normal distribution, t tests (df, 1439) comparing the observed values to those under the standard normal distribution were conducted at each false positive rate, for each of the indices. The significance level for each t test was set at .008 after the adjustment according to the Bonferroni procedure (i.e. .05/6). All of the values were significantly different from those expected under the normal distribution, except for ECIZ4 at the .10 false positive rate[3].

In sum these results indicate that over the different combinations of conditions, the cutoff values varied considerably. The ECIZ4 values tended to be closer than the Lz values to the tabled values of a standard normal distribution. On the whole, the cutoff values for both indices differed significantly from the tabled values of a standard normal distribution, but the tests were very powerful, each with 1439 degree of freedom. Over the different combinations of conditions there was a very slight tendency for Lz to be less extreme than the standard normal distribution and a tendency for ECIZ4 to be more extreme.

Effect of independent variables

To examine the effects of the four independent variables on the cutoff values for each of the indices, ANOVAS were conducted at each of the three cutoffs. The results of these ANOVAS are summarized in Table 7. The significance level was set at .008 after adjustment according to the Bonferroni procedure.

---

[3]It is noted that there is a non-independence within groups for the t tests which results in an inflation of Type I error and a reduction in power (Zimmerman & Zumbo, 1992). However, because the sample size was very large, power was not an issue in this study.

Table 7
Summary of ANOVA Results for Three False Positive Cutoff Values

SOURCE OF VARIATION	PROBABILITY FOR LZ			PROBABILITY FOR ECIZ4		
	.01	.05	.10	.01	.05	.10
Sampsiz	.000*	.000*	.000*	.000*	.000*	.000*
Variance of Diff (Diff)	.000*	.000*	.000*	.000*	.000*	.062
Discrimination (Disc)	.000*	.000*	.000*	.487	.000*	.000*
Model	.063	.000*	.000*	.000*	.000*	.000*
Sampsiz x Diff	.163	.289	.383	.111	.003*	.007*
Sampsiz x Disc	.605	.579	.970	.893	.176	.219
Sampsiz x Model	.757	.712	.584	.112	.555	.054
Diff x Disc	.002*	.721	.057	.000*	.000*	.000*
Diff x Model	.000*	.000*	.000*	.000*	.000*	.002*
Disc x Model	.366	.297	.396	.815	.505	.219
Sampsiz x Diff x Disc	.366	.297	.396	.815	.505	.219
Sampsiz x Diff x Model	.356	.530	.534	.238	.159	.297
Sampsiz x Disc x Model	.500	.262	.151	.149	.225	.423
Diff x Disc x Model	.000*	.000*	.001*	.000*	.000*	.025
Sampsiz x Diff x Disc x Model	.928	.191	.166	.091	.291	.260

*p <.008

A number of interactions occurred at the three false positive rates for both indices. A Diff x Model interaction occurred at all false positive rates for both indices. A three way Diff x Disc x Model interaction also occurred for both indices at all false positive rates, with the exception of ECIZ4 at the .10 false positive rate.

A three way interaction occurs when the interaction of two of the variables is not the same at all levels of the third variable (Keppel, 1991). When plotted, the three way interaction for Lz showed the same pattern for the cutoff values at all three false positive rates. This pattern is illustrated in Figure 2 for the cutoff value at the .01 false positive rate. For low Disc, the 3PLM produced more extreme cutoff values than the 2PLM at low Diff, but there was no difference at high Diff. For high Disc a disordinal Diff x Model interaction pattern was seen. The combination of the 2PLM and high Diff produced the most extreme values.

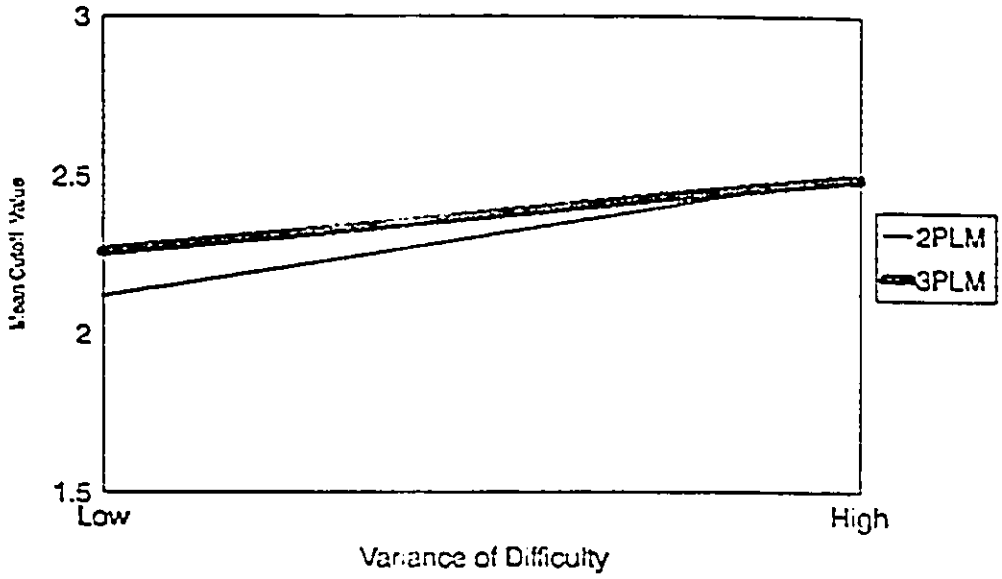
The pattern of the Diff x Disc x Model interaction for ECIZ4 was different than for Lz. The pattern, which was the same for the three false positive rates, is illustrated in Figure 3 for the cutoff values at the .01 false positive rate. This figure shows that, while the interaction of Diff x Model was ordinal at both levels of Disc, the slope of the 2PLM line was steeper at high Disc.

For ECIZ4, there was a Sampsiz x Diff interaction at the .05 and .10 false positive rates. The interaction was different at the two false positive rates. At the .05 false positive rate, low Sampsiz produced more extreme values than high Sampsiz at high Diff, but the effect was not significant for low Diff. At the .10 false positive rate, low Sampsiz produced more extreme values at both levels of Diff. However, high Diff produced more extreme values than low Diff for high Sampsiz, while high Diff produced less extreme values than low Diff for low Sampsiz. The main effect for Sampsiz for Lz occurred without any interactions. At all three false positive rates, the Sampsiz of 1000 produced more extreme values than the Sampsiz of 2500. The results of these analyses indicate that the independent

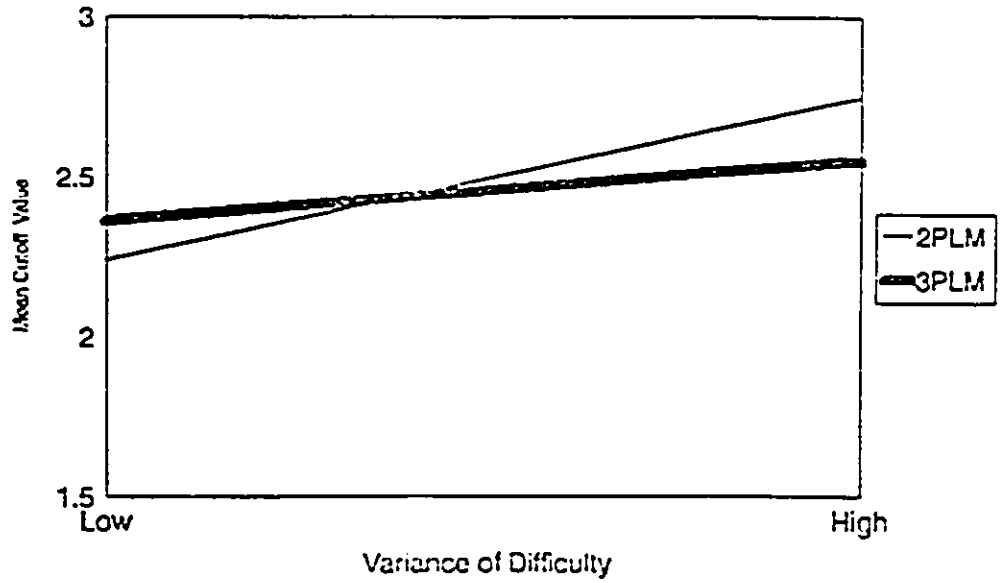
Figure 2

Diff x Disc x Model Interaction for Lz Cutoff

Values at the .01 False Positive Rate



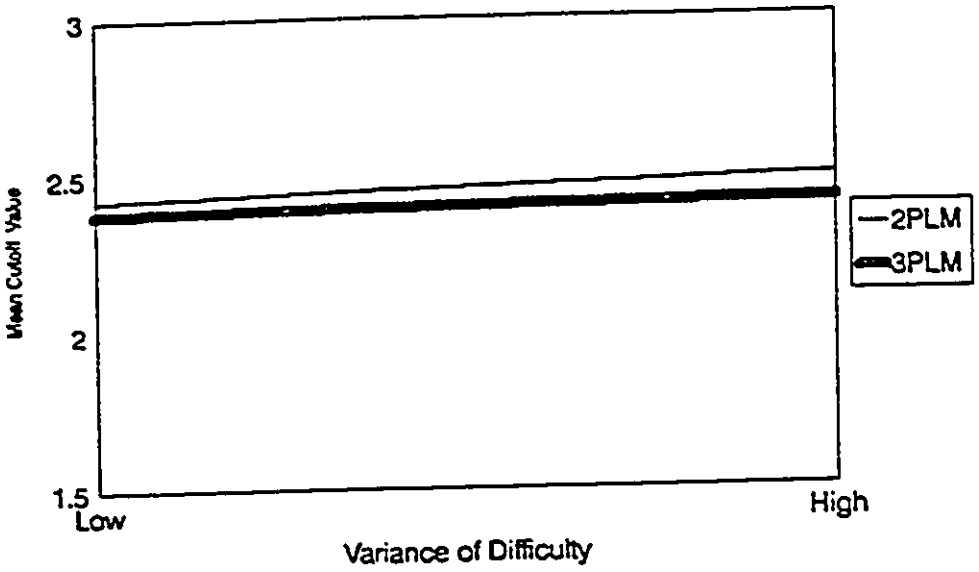
Low Disc



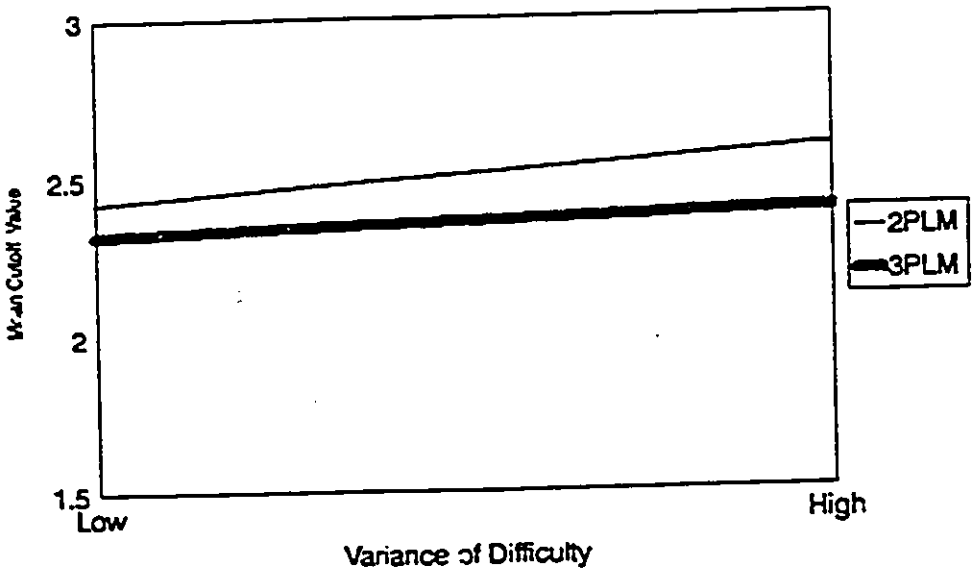
High Disc

Figure 3

Diff x Disc x Model Interaction for ECIZ4 Cutoff Values at the .01 False Positive Rate



Low Disc



High Disc

variables had a complex effect on the cutoff values. Furthermore, a general characterization is not possible.

Discussion

If cutoff values obtained from simulated nonaberrant response vectors are to be used in practical applications, it is desirable that they be stable under different test conditions. The cutoff values for the various combinations of independent variables ranged considerably. The range of the mean cutoff values over conditions was less for ECIZ4 than for Lz. Only one other study has produced Monte Carlo data on the cutoff values of the Lz and ECIZ4 indices. Noonan (1989) found that the cutoff values for ECIZ4 were less variable than for Lz when different test lengths and IRT models were used.

While the cutoff values varied considerably over conditions the practical impact on detection rates was not likely to be very large. This impact can be estimated by examining the differences in detection rates produced by using cutoff values from the nonaberrant sample and detection rates using cutoff values from a standard normal distribution (Tables 17 and 18). For example, differences in detection rates for the marginals never exceeded .04.

In practical test situations it is unlikely that the generation of nonaberrant response vectors will be replicated. Therefore, it is important to have some knowledge of the stability of the cutoff values over replications. As expected, the standard deviations of the cutoff values for the indices over the 90 replications were greater at the .01 false positive rate than at .05 or .10, indicating that the values are least stable in the tails of the distributions. The cutoff values for Lz were found to be approximately twice as variable over replications as for ECIZ4. The stability of the cutoff values for these two indices has only been examined in one other study, in which it was also found that ECIZ4 was more stable (Noonan, 1989).

The standard deviations of the cutoff values were no greater than .17. Generally these deviations would only produce moderate impacts on detection

rates. However, depending upon the replication a researcher might obtain decidedly different cutoff values. If a researcher were to simulate aberrance without any replications a chance error in the detection rates could occur which would be exacerbated by these deviations in the cutoff values. These errors could be quite large and could have important consequences, especially in a research setting.

These large standard deviations over replications are to be expected at the .01 false positive rate, given the sparse data in the extreme tails of the distribution. There are two possible approaches to reducing these deviations. One possibility that should be considered in future research is to smooth the distributions prior to setting the cutoff scores. Another approach to increase the number of response vectors per replication.

Several observations were made regarding the extremeness of the cutoff values. First it was noted that over the entire sample, the values for ECIZ4 were more extreme than for Lz. It was also clearly evident that over the entire sample low Sampsiz produced more extreme values, while high Diff produced more extreme values. The most extreme values were produced by the combination of high Diff, high Disc, and the 2PLM.

Data regarding the extremeness of the two indices is only available in one other study in which Noonan (1989) found that more extreme values were produced for Lz. However, Noonan's research did not involve manipulations of the Sampsiz, Diff, and Disc parameters. The difference in the observed values may also have resulted because of an important difference in the methodology for the two studies. In the present study, the values of the indices were calculated with item parameters that had been estimated using LOGIST, whereas in the Noonan study, the actual item parameters that had been used to generate the non-aberrant response vectors were used.

Table 8 can be used to compare observed cutoff values over the entire sample to those obtained by Noonan (1989) and those of the standard normal distribution. The observed values were less extreme than in the Noonan study. They were also closer to the standard normal, except for Lz at the

.10 false positive rate. As noted above, there were several important differences in the conditions and methodology used for the two studies. Differences in the Diff or Sampsiz variables cannot account for difference in the extremeness of the values for the two studies. The Diff variable was held constant in the Noonan study, in the range of -2.00 to 2.00. However, on average this is the same as for the present study in which half of the conditions had a Diff range of -2.5 to 2.5 while the other half had a Diff range of -1.5 to 1.5. In the Noonan study, the actual item parameters that were used in the generation of the items were also used for the calculation of the indices. It is possible that the use of estimated item parameters for the calculation of the results may have made some difference. The program that generated the response vectors uses a probability for determining if the response to an item is correct or not. Estimated parameters may more accurately represent the response vectors than the actual parameters.

As compared to the Noonan data, the values for the present study are likely more useful for practical applications. First, the observed values are likely more generalizable because a larger number of independent variables was manipulated. Second, in practical testing situations estimated item parameters are used.

Most of the previous research on the distributions of non-aberrant appropriateness indices has been concentrated on the mean and standard deviation of the distribution and in most of the research the conclusion has been that the distributions while close to normal do not follow a normal distribution exactly (Drasgow, Levine, & Mclaughlin, 1987; Birenbaum, 1985; Harnisch & Tatsuoka, 1983; Noonan, 1989). Very little research has been conducted on the distribution of the indices in the tails of the distribution, but in general the research has suggested that the tails of the distributions do not correspond closely to the values expected under a standard normal distribution (Drasgow, Levine, & Williams, 1985; Drasgow & Guertler, 1987; Noonan, 1989).

Table 8

Comparison of Cutoff Values at Three False Positive Rates

	Cole	Noonan	Standard normal
.01 False Positive			
Lz	- 2.41	- 2.69	- 2.33
ZCIZ4	2.43	2.48	2.33
.05 False Positive			
Lz	- 1.61	- 1.79	- 1.65
ZCIZ4	1.67	1.71	1.65
.10 False Positive			
Lz	- 1.21	- 1.34	- 1.28
ZCIZ4	1.28	1.32	1.28

The observed cutoff values were compared to the cutoff values for the standard normal distribution for each index at each false positive rate. Five out of the six t-tests comparing the cutoff values for the indices to those of the standard normal distribution were significant. Therefore, strictly speaking the cutoff values of the indices do not correspond to those of a standard normal distribution. However, a very large sample size was used for each of the t - tests so very small differences resulted in statistical significance.

In simulation research in which nonaberrant samples were simulated, Lz was found to be approximately normal N(0,1) by Drasgow, Levine, and Williams (1985), while Tatsuoka and Tatsuoka (1982) considered ZCIZ4 to be approximately normal. Noonan (1989) found that while both indices had means and standard deviations that approximated a N(0,1) distribution, the distributions were skewed, indicating non-normality in the tails of the distribution that were used to identify aberrance.

Several researchers have suggested that the distribution of appropriateness indices in non-aberrant conditions may be unique for different sets of test conditions (Smith, 1982; Molenaar & Hoijtink, 1990; Noonan, 1989). At least two researchers, Gafni (1987) and Anderson (1990), have suggested using simulations of the test under non-aberrant conditions

for establishing cutoff values. The results of the ANOVAS for the effects of the independent variables on cutoff values suggest that while the IVs have an effect on the cutoff values, the effects are quite complex and difficult to characterize. This implies that for every test it may be necessary to simulate non-aberrant response vectors to obtain accurate cutoff values. However, whether this procedure is necessary or whether cutoff values from the standard normal distribution provide a reasonable proxy is a question which is addressed in the third section of this chapter, in which detection rates using observed cutoff values are compared to detection rates using cutoff values from the standard normal distribution.

Detection Rates

The detection rates of Lz and ECIZ4 were computed for both high and low aberrance for the 16 combinations of Diff, Disc, Model, and Sampsiz. In this section the effects of the independent variables on detection rates are examined. It is also in this section that the stability of the detection rates, the detection rates for the two indices, as well as detection rates for high and low aberrance are discussed.

In order to determine the effects of the independent variables on detection rates, for each index six ANOVAS were conducted (one at each of the three false positive rates for high aberrance and one at each of the false positive rates for low aberrance). The results for each of the ANOVAS for each index and for each type of aberrance is examined. An overall summary of these results is then presented and discussed.

As mentioned earlier, the significance tests for the ANOVAS were conducted at the .05 level, but the Bonferroni procedure was used so as to avoid capitalizing on chance. Since there were 12 ANOVAS overall, the adjusted significance level for each ANOVA was .004 (i.e. .05/12).

The procedure for the interpretation of the results of each of the ANOVAS was the same throughout. Overall there were several two way interactions for the ANOVAS, but no higher order interactions. Where

significant interactions were present they were plotted and four t tests were conducted to test for simple effects. The significance level for each of the t tests was set at .05 and was not adjusted for the possible increase in familywise type I error. This decision is controversial and is based on two reasons cited by Keppel (1991). First, some form of control operates when we restrict ourselves to the further analyses of only significant effects. Second, type I errors may readily be discovered as researchers attempt to replicate significant and interesting findings.

High aberrance for Lz

The detection rates for Lz for high aberrance are shown in Table 9. Over the entire sample the mean detection rates for Lz were .498, .638, and .707 for the .01, .05, and .10 false positive rates, respectively. The detection rates varied considerably over combinations of conditions. At the .01 false positive rate, detection rates ranged from .286 to .782, a spread of .496. The range of the detection rates at the .05 and .10 false positive rates was .459 and .420 respectively. Detection rates also varied considerably over replications. For example, the lowest standard deviation of the detection rates at the .01 false positive rate was .073, which occurred for a detection rate of .286. Thus, if detection rates were normally distributed, for this example 32% of the replications would have had detection rates that were below .213 or above .359.

The marginals in Table 9 reveal interesting differences in the detection rates. The detection rates for high Diff were considerably higher than for low Diff. The difference was .29, .27, and .24 at the .01, .05, and .10 false positive rates, respectively. Similarly, the detection rates for the 2PLM were considerably higher than for the 3PLM. The difference was .15, .16, and .16 at the .01, .05, and .10 false positive rates, respectively. However, the difference for the detection rates for Disc were moderate and for Sampsiz they were not very high. High Disc was more effective than low Disc (a difference of .06, .04, and .02 at the .01, .05, and .10 false

Table 9

Mean and Standard Deviation of Detection Rates for High Aberrant Response Patterns for Lz at Three False Positive Rates Over 90 Replications

CONDITION				FALSE POSITIVE RATE					
				.01		.05		.10	
Sampsiz	Variance of Diff	Disc	Model	Mean	SD	Mean	SD	Mean	SD
1000	low	low	2PLM	.330	.089	.517	.088	.627	.083
1000	low	low	3PLM	.289	.083	.438	.088	.518	.086
1000	low	high	2PLM	.425	.091	.573	.086	.656	.080
1000	low	high	3PLM	.371	.078	.486	.072	.546	.069
1000	high	low	2PLM	.761	.104	.888	.066	.933	.046
1000	high	low	3PLM	.492	.129	.643	.123	.714	.115
1000	high	high	2PLM	.774	.110	.892	.066	.936	.045
1000	high	high	3PLM	.569	.140	.696	.121	.755	.108
2500	low	low	2PLM	.352	.091	.532	.091	.638	.084
2500	low	low	3PLM	.286	.073	.437	.076	.518	.074
2500	low	high	2PLM	.420	.098	.571	.089	.655	.083
2500	low	high	3PLM	.355	.087	.469	.082	.528	.080
2500	high	low	2PLM	.725	.114	.865	.078	.916	.058
2500	high	low	3PLM	.497	.123	.642	.107	.715	.098
2500	high	high	2PLM	.782	.111	.896	.070	.938	.048
2500	high	high	3PLM	.534	.129	.661	.120	.723	.112
Entire Sample				.498	.201	.638	.185	.707	.169
MARGINALS									
Sampsiz	1000			.50	.20	.64	.19	.71	.17
	2500			.49	.20	.63	.17	.70	.18
Variance of Diff.	low			.35	.10	.50	.10	.59	.10
	high			.64	.17	.77	.15	.83	.13
Disc.	low			.47	.20	.62	.19	.70	.17
	high			.53	.19	.66	.18	.72	.17
Model	2PLM			.57	.22	.72	.19	.79	.16
	3PLM			.42	.15	.56	.14	.63	.14

positive rates, respectively) and low Sampsiz was more effective than high Sampsiz (a difference of .01 at all three false positive rates).

The results of the ANOVAS for Lz for high aberrance are presented in Table 10. The effects for Diff and Model were significant in the presence of a significant Diff x Model interaction, for all three false positive rates. The Diff x Model interaction was ordinal and for all comparisons differences were found which supported the findings of main effects. The main effect for Diff occurred because high Diff was more effective than low Diff. The main effect for Model occurred because the 2PLM was more effective than the 3PLM. A comparison of the mean difference between the detection rates for the combinations of variables showed that the combination of the 2PLM and high Diff enhanced the detection of aberrance.

The Diff x Disc interaction was also significant, but only at the .01 false positive rate. For this false positive rate, the main effects of Diff and Disc were also significant. The Diff x Disc interaction was ordinal and for all comparisons differences were found which supported the findings of main effects. The main effect for Diff and Disc occurred because high Diff was more effective than low Diff and high Disc was more effective than low Disc. A comparison of the mean difference between the detection rates for the combinations of variables showed that the combination of low Diff and high Disc enhanced the detection of aberrance.

At the .05 and .10 false positive rates, the main effect of Disc was also significant, with no interactions involving Disc. Again high Disc was more effective than low Disc.

Low aberrance for Lz

The detection rates for Lz, for low aberrance are shown in Table 11 . The detection rates for low aberrance were .583, .676, and .730 for the .01, .05, and .10 false positive rates, respectively. These detection rates were higher than for high aberrance. The difference was greatest at the .01 false positive rate, 8%, and decreased to 4% at the .05 false positive rate and 2% at the .10 false positive rate. As for high aberrance, the detection rates

Table 10

Summary of ANOVA Results for Lz High and Low Aberrance

df = (1,1424)

	PROBABILITY FOR HIGH ABERRANCE			PROBABILITY FOR LOW ABERRANCE		
SOURCE OF VARIATION	FALSE POSITIVE RATE			FALSE POSITIVE RATE		
	.01	.05	.10	.01	.05	.10
Sample Size (Sampsiz)	.172	.117	.121	.673	.704	.580
Variance of Diff (Diff)	.000*	.000*	.000*	.000*	.000*	.000*
Discrimination (Disc)	.000*	.000*	.000*	.000*	.000*	.000*
Model	.000*	.000*	.000*	.001*	.004*	.030
Sampsiz x Diff	.209	.200	.283	.578	.981	.810
Sampsiz x Disc	.422	.283	.194	.017	.035	.094
Sampsiz x Model	.395	.229	.231	.904	.990	.315
Diff x Disc	.003*	.079	.776	.000*	.015	.007
Diff x Model	.000*	.000*	.000*	.620	.018	.012
Disc x Model	.473	.570	.648	.631	.242	.280
Sampsiz x Diff x Disc	.313	.491	.658	.534	.387	.529
Sampsiz x Diff x Model	.456	.692	.787	.845	.841	.969
Sampsiz x Disc x Model	.102	.117	.091	.014	.052	.084
Diff x Disc x Model	.203	.183	.361	.038	.025	.149
Sampsiz x Disc x Disc x Model	.027	.096	.175	.730	.487	.970

*p < .004

Table 11

Mean and Standard Deviation of Detection Rates for Low Aberrant Response Patterns for Lz at Three False Positive Rates Over 90 Replications

CONDITION				FALSE POSITIVE RATE					
				.01		.05		.10	
Sampsiz	Diff	Disc	Model	Mean	SD	Mean	SD	Mean	SD
1000	low	low	2PLM	.308	.087	.470	.083	.555	.076
1000	low	low	3PLM	.314	.106	.456	.101	.534	.095
1000	low	high	2PLM	.465	.083	.564	.103	.649	.153
1000	low	high	3PLM	.433	.078	.526	.088	.604	.139
1000	high	low	2PLM	.751	.112	.820	.099	.854	.091
1000	high	low	3PLM	.732	.122	.812	.105	.846	.096
1000	high	high	2PLM	.836	.098	.882	.098	.908	.093
1000	high	high	3PLM	.815	.112	.882	.108	.908	.104
2500	low	low	2PLM	.329	.088	.482	.087	.563	.082
2500	low	low	3PLM	.312	.102	.453	.095	.534	.089
2500	low	high	2PLM	.446	.080	.551	.075	.625	.151
2500	low	high	3PLM	.432	.076	.524	.075	.612	.146
2500	high	low	2PLM	.787	.092	.846	.083	.867	.081
2500	high	low	3PLM	.739	.115	.813	.098	.850	.094
2500	high	high	2PLM	.811	.113	.850	.113	.874	.115
2500	high	high	3PLM	.819	.111	.879	.111	.906	.105
Entire Sample				.583	.232	.676	.201	.730	.186
MARGINALS									
Sampsiz	1000			.58	.23	.68	.20	.73	.19
	2500			.58	.23	.67	.20	.73	.19
Variance of Diff	low			.38	.11	.50	.10	.58	.13
	high			.79	.12	.85	.11	.88	.10
Disc	low			.53	.21	.64	.19	.70	.19
	high			.63	.24	.71	.20	.76	.18
Model	2PLM			.59	.23	.68	.19	.74	.18
	3PLM			.57	.23	.67	.21	.72	.19

varied considerably over combinations of conditions and the spread was also similar to the spread for high aberrance. At the .01 false positive rate, detection rates ranged from .308 to .836, a spread of .528. The range of the detection rates at the .05 and .10 false positive rates was .429 and .374 respectively. As for high aberrance detection rates varied considerably over replications. For example, the lowest standard deviation of the detection rates at the .01 false positive rate was .076, which occurred for a detection rate of .432. Thus, if detection rates were normally distributed, for this example 32% of the replications would have had detection rates that were below .356 or above .508.

The marginals in Table 11 reveal interesting differences in the detection rates. As for high aberrance, the detection rates for high Diff were considerably higher than for low Diff. The difference was .41, .35, and .30 at the .01, .05, and .10 false positive rates, respectively. However, the differences were not as high for the other three independent variables. The detection rates for the 2PLM were higher than for the 3PLM, but the difference was only .02 .01, and .02 at the .01, .05, and .10 false positive rates, respectively. As for high aberrance, high Disc was moderately more effective than low Disc (a difference of .10, .07, and .06 at the .01, .05, and .10 false positive rates, respectively). There was almost no difference for the two levels of Sampsiz at any of the three false positive rates.

The results of the ANOVAS for Lz for low aberrance are presented in Table 10. At the .01 false positive rate the effects for Diff, Disc, and Model were significant. The only interaction which was significant was for Diff x Disc at the .01 false positive rate. The Diff x Disc interaction was ordinal and for all comparisons, differences were found which supported the findings of main effects. The interaction was significant because the combination of low Diff and high Disc enhanced the detection of aberrance. As for high aberrance, the main effect for Diff occurred because high Diff was more effective than low Diff; the main effect for Disc occurred because

Table 12

Mean and Standard Deviation of Detection Rates for High Aberrant Response Patterns for ECIZ4 at Three False Positive Rates Over 90 Replications

CONDITION				FALSE POSITIVE RATE					
				.01		.05		.10	
Sampsiz	Variance of Diff	Disc	Model	Mean	SD	Mean	SD	Mean	SD
1000	low	low	2PLM	.422	.104	.653	.091	.765	.073
1000	low	low	3PLM	.308	.096	.514	.101	.636	.092
1000	low	high	2PLM	.455	.100	.651	.087	.744	.076
1000	low	high	3PLM	.349	.083	.552	.091	.675	.088
1000	high	low	2PLM	.768	.099	.906	.056	.948	.035
1000	high	low	3PLM	.375	.134	.603	.135	.713	.121
1000	high	high	2PLM	.728	.125	.885	.072	.937	.045
1000	high	high	3PLM	.424	.152	.639	.132	.744	.110
2500	low	low	2PLM	.437	.110	.664	.094	.769	.076
2500	low	low	3PLM	.294	.086	.495	.090	.609	.081
2500	low	high	2PLM	.457	.108	.649	.092	.746	.078
2500	low	high	3PLM	.317	.096	.504	.097	.616	.093
2500	high	low	2PLM	.732	.109	.886	.064	.936	.043
2500	high	low	3PLM	.362	.127	.581	.121	.695	.105
2500	high	high	2PLM	.744	.120	.890	.072	.939	.045
2500	high	high	3PLM	.327	.141	.540	.142	.653	.131
Entire Sample				.469	.201	.663	.173	.758	.144
MARGINALS									
Sampsiz	1000			.48	.20	.68	.17	.77	.19
	2500			.46	.20	.65	.18	.75	.15
Variance of Diff.	low			.38	.12	.59	.12	.70	.10
	high			.56	.23	.74	.18	.82	.15
Disc.	low			.46	.20	.66	.17	.76	.14
	high			.48	.20	.66	.17	.76	.14
Model	2PLM			.59	.19	.77	.14	.85	.11
	3PLM			.34	.12	.55	.12	.67	.11

high Disc was more effective than low Disc. As for high aberrance, the main effect for Model occurred because the 2PLM was more effective than the 3PLM.

At the .05 false positive rate, the main effects of Diff, Disc, and Model were significant and there were no interactions. Again, high Diff was more effective than low Diff, high Disc was more effective than low Disc and the 2PLM was more effective than the 3PLM.

At the .10 false positive rate, there were no interactions, but the main effects of Diff and Disc were significant. As above, high Diff was more effective than low Diff and high Disc was more effective that low Disc.

High aberrance for ECIZ4

The detection rates for ECIZ4 for high aberrance are presented in Table 12. The overall detection rates for high aberrance were .469, .663, and .758 for the .01, .05, and .10 false positive rates, respectively. Compared to the detection rates obtained for Lz for high aberrance, these detection rates were 3% lower at the .01 false positive rate, but 2% and 5% higher at the .05 and .10 false positive rates, respectively. As for Lz for high and low aberrance, the detection rates varied considerably over combinations of conditions. At the .01 false positive rate, detection rates ranged from .297 to .768, a spread of .474. The range of the detection rates at the .05 and .10 false positive rates was .411 and .330, respectively. This spread in detection rates over the conditions was slightly lower than the spread for high aberrance for Lz and slightly lower than the spread for low aberrance for Lz. As for Lz, detection rates varied considerably over replications. For example, the lowest standard deviation of the detection rates at the .01 false positive rate was .083, which occurred for a detection rate of .349. Thus, if detection rates were normally distributed, for this example, 32% of the replications would have had detection rates that were below .266 or above .432.

The marginals in Table 12 show the same pattern of differences as for Lz for high aberrance. The detection rates for high Diff were considerably higher than for low Diff and the detection rates for the 2PLM were

considerably higher than for the 3PLM. The difference for the Diff condition was .18, .15, and .12 at the .01, .05, and .10 false positive rates, respectively. The difference for the Model variable was .25, .22, and .18 at the .01, .05, and .10 false positive rates, respectively. However, there was very little difference for the detection rates for Disc and Sampsiz. High Disc was more effective than low Disc (a difference of .02, .00, and .00 at the .01, .05, and .10 false positive rates, respectively) and low Sampsiz was more effective than high Sampsiz (a difference of .02, .03, and .02 at the .01, .05, and .10 false positive rates, respectively).

The results of the ANOVAS were the same for all three false positive rates (see Table 13). Sampsiz, Diff, and Model were significant in the presence of Sampsiz x Model and Diff x Model interactions.

The analyses of the interactions for simple effects revealed consistent results for the three false positive rates. As for Lz for high aberrance, the Diff x Model interaction was ordinal and for all comparisons, differences were found which supported the findings of main effects. The interaction of Diff x Model was significant because the combination of the 2PLM and high Diff enhanced the detection of aberrance. Again, the main effects for Model and Diff occurred because the 2PLM was more effective than the 3PLM and high Diff produced better detection rates than low Diff.

The post - hoc analysis of the Sampsiz x Model interaction showed that low Sampsiz was more effective than high Sampsiz for the 3PLM only.

Low aberrance for ECIZ4

The detection rates for ECIZ4 for low aberrance are presented in Table 14. The overall detection rates were .625, .741, and .830 for the .01, .05, and .10 false positive rates, respectively. These detection rates were higher than those obtained for Lz for low aberrance (4%, 4%, and 10% higher at the .01, .05, and .10 false positive rates, respectively). As occurred for Lz, for the entire sample the detection rates for low aberrance were higher than the detection rates for high aberrance. The difference was greatest at the

Table 13

Summary of ANOVA Results for ECIZ4 for High and Low Aberrance

df = (1,1424)

	PROBABILITY FOR HIGH ABERRANCE			PROBABILITY FOR LOW ABERRANCE		
SOURCE OF VARIATION	FALSE POSITIVE RATE			FALSE POSITIVE RATE		
	.01	.05	.10	.01	.05	.10
Sample Size (Sampsiz)	.001*	.000*	.000*	.473	.738	.104
Variance of Diff (Diff)	.000*	.000*	.000*	.000*	.000*	.000*
Discrimination (Disc)	.032	.852	.658	.000*	.007	.096
Model	.000*	.000*	.000*	.000*	.000*	.000*
Sampsiz x Diff	.033	.065	.279	.878	.633	.140
Sampsiz x Disc	.179	.027	.010	.076	.038	.576
Sampsiz x Model	.001*	.000*	.000*	.971	.387	.376
Diff x Disc	.006	.214	.596	.030	.011	.469
Diff x Model	.000*	.000*	.000*	.000*	.000*	.000*
Disc x Model	.276	.066	.018	.000*	.059	.000*
Sampsiz x Diff x Disc	.962	.790	.491	.446	.168	.534
Sampsiz x Diff x Model	.570	.446	.309	.295	.311	.332
Sampsiz x Disc x Model	.009	.009	.011	.662	.431	.354
Diff x Disc x Model	.507	.221	.012	.022	.115	.179
Sampsiz x Disc x Disc x Model	.010	.041	.106	.850	.577	.950

*p < .004

Table 14

Mean and Standard Deviation of Detection Rates for Low Aberrant Response Patterns for ECIZ4 at Three False Positive Rates Over 90 Replications

CONDITION				FALSE POSITIVE RATE					
				.01		.05		.10	
Sampsiz	Variance of Diff.	Disc.	Model	Mean	SD	Mean	SD	Mean	SD
1000	low	low	2PLM	.449	.087	.625	.106	.799	.150
1000	low	low	3PLM	.406	.100	.558	.088	.640	.100
1000	low	high	2PLM	.505	.080	.621	.119	.826	.179
1000	low	high	3PLM	.438	.073	.539	.067	.601	.103
1000	high	low	2PLM	.851	.115	.972	.039	.989	.021
1000	high	low	3PLM	.684	.127	.810	.108	.877	.103
1000	high	high	2PLM	.943	.082	.995	.009	.998	.004
1000	high	high	3PLM	.708	.117	.799	.100	.864	.106
2500	low	low	2PLM	.465	.093	.642	.119	.801	.147
2500	low	low	3PLM	.425	.085	.580	.074	.674	.100
2500	low	high	2PLM	.488	.071	.594	.109	.839	.181
2500	low	high	3PLM	.437	.071	.542	.065	.626	.134
2500	high	low	2PLM	.866	.095	.976	.029	.991	.012
2500	high	low	3PLM	.685	.121	.811	.091	.901	.079
2500	high	high	2PLM	.945	.083	.990	.034	.998	.004
2500	high	high	3PLM	.702	.129	.795	.114	.852	.109
Entire Sample				.625	.214	.741	.188	.830	.171
MARGINALS									
Sampsiz	1000			.62	.21	.74	.19	.83	.17
	2500			.63	.21	.74	.19	.84	.17
Variance of Diff.	low			.45	.09	.59	.10	.73	.17
	high			.80	.15	.89	.12	.94	.09
Disc.	low			.60	.21	.75	.18	.84	.16
	high			.65	.22	.73	.20	.83	.18
Model	2PLM			.69	.23	.80	.20	.91	.15
	3PLM			.56	.17	.68	.15	.76	.16

.01 false positive rate, 16%, 8% at the .05 false positive rate, and 7% at the .10 false positive rate. Again, the detection rates varied considerably over combinations of conditions. At the .01 false positive rate, detection rates ranged from .406 to .945, a spread of .539. The spread of the detection rates at the .05 and .10 false positive rates was .456 and .397, respectively. This spread was similar to the spreads observed for Lz and for ECIZ for high aberrance. As for previous observations, detection rates varied considerably over replications. For example, the lowest standard deviation of the detection rates at the .01 false positive rate was .071, which occurred for two detection rates (.437 and .488).

The marginals in Table 14 show the same pattern of differences as for Lz for low aberrance. The detection rates for high Diff were considerably higher than for low Diff. The difference was .35, .30, and .31 at the .01, .05, and .10 false positive rates, respectively. The differences for the other three independent variables were not very high. The detection rates for the 2PLM were higher than for the 3PLM, but the differences were not high as for high aberrance. The difference was .07, .12, and .15 at the .01, .05, and .10 false positive rates, respectively. There were almost no differences for Sampsiz and the differences for Disc were small. High Disc was more effective than low Disc (a difference of .05, .02, and .01 at the .01, .05, and .10 false positive rates, respectively).

The results of the ANOVAS for ECIZ4 for low aberrance are summarized in Table 13. At the .01 false positive rate, significant effects were found for Diff, Disc, and Model in the presence of Diff x Model and Disc x Model interactions. The Diff x Model interaction was ordinal and for all comparisons, differences were found which supported the findings of main effects. The interaction of Diff x Model was significant because the combination of the 2PLM and high Diff enhanced the detection of aberrance. The analysis of the ordinal interaction of Disc x Model for simple effects showed that high Disc was more effective than low Disc, but only for the 3PLM. As had been found in the previous analyses, the main effects for Model and Diff

occurred because the 2PLM was more effective than the 3PLM and high Diff produced better detection rates than low Diff.

At the .05 false positive rate, the effects for Diff and Model were significant in the presence of an ordinal Diff x Model interaction. As for the .01 false positive rate, the combination of the 2PLM and high Diff produced enhanced detection rates. The significant main effects for Model and Diff were present because the 2PLM was more effective than the 3PLM, while high Diff was more effective than low Diff.

At the .10 false positive rate, the effects for Diff and Model were significant in the presence of Diff x Model and Disc x Model interactions. The Diff x Model interaction was ordinal. The 2PLM was more effective than the 3PLM, while high Diff was more effective than low Diff. At this false positive rate the combination of the 2PLM and low Diff enhanced detection. The Disc x Model interaction was also ordinal, but a simple main effect was found for Disc. Low Disc was more effective, but only for the 3PLM.

Discussion

The overall detection rates were quite high. The detection rates for low aberrance by Lz were 58%, 68%, and 73% for the .01, .05, and .10 false positive rates, respectively. The corresponding detection rates for low aberrance for ECIZ4 were 63%, 74%, and 83%. For high aberrance the detection rates by Lz were 50%, 64%, and 71% for the .01, .05, and .10 false positive rates, respectively. The corresponding rates for high aberrance for ECIZ4 were 47%, 66%, and 76%.

It is difficult to compare detection rates in this study with those in other studies because of differences in the test conditions and methodology used. Longer tests have been found to produce better detection rates than shorter tests and higher levels of aberrance have been found to result in better detection rates than lower levels. The technique used to generate aberrance can also influence the detection rates. For a given percentage of aberrance, the Rudner technique should result in higher detection rates than the Levine technique. This is because for a given percentage of aberrance a

greater number of responses are changed with the Rudner technique. A number of studies have shown that detection rates improve as the ability levels become more extreme (Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, Williams, McLaughlin, & Candell, 1987; Gafni, 1987; Noonan, 1989). In this study high aberrance was generated for response vectors in the -3.00 to 0 ability range, while low aberrance was generated for response vectors in the 0 to 3.00 ability range. Within these ranges, a uniform distribution of ability was used.

In the present study, a 60 item test was used and 16.6% aberrance was generated using the Rudner technique. The ability range used to generate aberrant responses was moderately extreme and within that range a uniform distribution of ability was used. In as much as it is possible to judge, the detection rates obtained in this study correspond to what might be expected given the results of previous research. For example, for Lz, Drasgow, Levine, and McLaughlin (1987) obtained detection rates of 87% for low aberrance generated from response vectors in the 93rd or higher percentile of ability on the 85 item SAT-V test. The Levine technique was used to generate aberrance and the 3PLM was used. This result was for 15% aberrance at the .05 false positive rate the detection. For the current study, the detection rate of low aberrance for Lz was 67% for the 3PLM at the .05 false positive rate for response vectors generated in the 50ieth or higher percentile.

Drasgow, Levine, McLaughlin, and Candell (1987) used the Levine technique to generate 17% aberrance on a 30 item test. For high aberrance generated from response vectors in the 0 the 9th percentiles, the detection rate for Lz using the 3PLM, was 33% at the .05 false positive rate. For low aberrance generated in the 93rd to 99th percentiles, the detection rate was 51%. The corresponding detection rates for Lz for the current study were 67% and 56%, respectively.

Noonan (1989), using the Rudner technique, generated 15% spuriously high and 15% spuriously low aberrance from response vectors in a normal ability range. A 40 item and an 80 item test were used and the results were

reported over the two tests. For the 3PLM, the detection rate at the .05 false positive rate was 22% for the high aberrance and 36% for the spuriously low sample.

For both indices, for high and low aberrance, detection rates varied considerably over replications. The variability over replications was greater for low aberrance than for high aberrance, but this could be expected since the detection rates for low aberrance were higher. The standard deviation of the mean detection rates for the 16 combination was consistently lower for ECIZ4 than for Lz. A comparison of the stability of the detection rates within combinations of conditions showed that the two indices had about equal stability over replications for low aberrance. For low aberrance, for the 16 combinations of conditions at the three false positive rates, the standard deviation of the mean detection rates for Lz was less than the standard deviation for ECIZ4 for 48% of the cases. However, for high aberrance, Lz tended to be more stable over replications. For the 16 combinations of conditions, the standard deviation for Lz was smaller than the standard deviation for ECIZ4 for 77% of the cases.

The variability of appropriateness indices over replications has never been examined in other research. In almost all simulation research on the detection rates of appropriateness indices no replications were performed. The results of this study showed that detection rates can vary considerably over replications. Therefore, to avoid chance results it would be prudent for future researchers to replicate their conditions.

For both indices, detection rates also varied considerably, and about equally, over different combinations of conditions for both high and low aberrance. The range was approximately .50, .45, and .40 at the .01, .05, and .10 false positive rates, respectively. This is as a result of the effects of the independent variables on detection rates which is discussed below.

Results of this study indicate that ECIZ4 generally had detection rates that were the same or better than Lz, but the difference for the overall sample was not great. The advantage to ECIZ4 increased as the false positive

rate increased. For low aberrance the difference ranged from 5% to 10%. For high aberrance the difference ranged from -3% to 5%.

While the performance of Lz and ECIZ4 have been compared in a number of studies, Noonan (1989) was the only one to note any systematic difference in the performance of the two indices. Noonan examined the performance of the two indices at the .01, .05, .10, and .25 false positive rates and noted that Lz produced higher rates of detection for lower false positive rates and lower levels of aberrance, while ECIZ4 produced higher rates of detection for higher false positive rates and higher levels of aberrance. Noonan concluded that the difference in the detection rates was not large.

The results support Noonan's finding that ECIZ4 produced higher levels of detection than Lz for higher false positive rates. The performance of Lz relative to ECIZ4 improved at lower false positive rates and for high aberrance at the .01 false positive rate Lz outperformed ECIZ4 (the difference was 3%). These results are congruent with Noonan's if one concludes that Lz improves relative to ECIZ4 as conditions make it more difficult to detect aberrance. It is more difficult to detect aberrance for lower false positive rates. It is also more difficult to detect aberrance for shorter tests and for lower levels of aberrance. In Noonan's research a 40 and 80 item test were used with three levels of aberrance. A review of Noonan's data revealed a tendancy for the relative performance of Lz to improve for the shorter test and lower levels of aberrance.

Over the entire sample, detection rates were higher for low aberrance than for high aberrance. The differences in detection rates for high and low aberrance are shown in Table 15 for the two indices. Over the entire sample the difference ranged from 2% to 8% for Lz and from 7% to 16% for ECIZ4. The difference was highest for high Diff and for the 3PLM. This is reflected in the differences over conditions. The greatest differences, favoring low aberrance, occurred for the conditions having high Diff and the 3PLM.

Previous studies in which comparisons were made for the detection rates of high and low aberrance have not given consistent results. However, in

Table 15
Difference Between Detection Rates for Low and High Aberrance for Lz and ECIZ4 at Three False Positive Rates

CONDITION				Lz			ECIZ4		
Sampsiz	Diff	Disc	Model	.01	.05	.01	.01	.05	.10
1000	low	low	2PLM	(0.022)	(0.047)	(0.072)	0.027	(0.028)	0.034
1000	low	low	3PLM	0.025	0.018	0.016	0.098	0.044	0.004
1000	low	high	2PLM	0.040	(0.009)	(0.007)	0.050	(0.030)	0.082
1000	low	high	3PLM	0.062	0.040	0.058	0.089	(0.013)	(0.074)
1000	high	low	2PLM	(0.010)	(0.068)	(0.088)	0.083	0.066	0.041
1000	high	low	3PLM	0.240	0.169	0.132	0.309	0.207	0.164
1000	high	high	2PLM	0.062	(0.010)	(0.028)	0.215	0.110	0.061
1000	high	high	3PLM	0.246	0.186	0.153	0.284	0.160	0.120
2500	low	low	2PLM	(0.023)	(0.050)	(0.075)	0.028	(0.022)	0.032
2500	low	low	3PLM	0.026	0.016	0.016	0.131	0.085	0.065
2500	low	high	2PLM	0.026	(0.020)	(0.030)	0.031	(0.055)	0.093
2500	low	high	3PLM	0.077	0.055	0.084	0.120	0.038	0.010
2500	high	low	2PLM	0.062	(0.019)	(0.049)	0.134	0.090	0.055
2500	high	low	3PLM	0.242	0.171	0.135	0.323	0.230	0.206
2500	high	high	2PLM	0.029	(0.046)	(0.064)	0.201	0.100	0.059
2500	high	high	3PLM	0.385	0.218	0.183	0.375	0.255	0.199
Entire Sample				0.085	0.038	0.023	0.156	0.078	0.072
MARGINALS									
Sampsiz	1000			0.080	0.040	0.020	0.140	0.060	0.060
	2500			0.090	0.040	0.030	0.170	0.090	0.090
Diff	low			0.030	0.000	(0.010)	0.070	0.000	0.030
	high			0.150	0.080	0.050	0.240	0.150	0.120
Disc	low			0.060	0.020	0.000	0.140	0.090	0.080
	high			0.100	0.050	0.040	0.170	0.070	0.070
Model	2PLM			0.020	(0.040)	(0.050)	0.100	0.030	0.060
	3PLM			0.150	0.110	0.090	0.220	0.130	0.090

these studies different item characteristics and ability ranges were used for generating the response vectors. The reason that low aberrance was better detected in this study may in part have resulted because there is less error in the high scores. A departure from the expected response is more notable because little guessing occurs for high ability levels. However, it appears that the item characteristics may play an important, and perhaps complex, role in determining the effectiveness of the indices for high and low aberrance.

Table 16 presents a summary of the above ANOVA results for the effects of the independent variables on detection rates. The main effects for the independent variables are shown at the top of the table. A "B" indicates the condition that produced better detection rates. When there is no entry (e.g., for Sampsiz) it means that there was no significant effect. The next listing in the table is for simple main effects. The last listing is for enhancements. "EN" indicates which combination of conditions produced enhanced detection rates.

Table 16 reveals a number of consistent results. High Diff was more effective than low Diff for both indices, for both high and low aberrance. This implies that for both indices and for both types of aberrance, the broader the range in the difficulty parameter for the items in a test, the higher the detection rate of aberrance.

The effect of the Diff variable on detection rates had been expected. In the only other research examining the effect of Diff on detection rates the same effect had been found, though a different methodology for generating aberrance had been used (Reise & Due, 1990). With a broader range in the difficulty parameter of the items there is more precision in estimating ability at the extremes of ability. Also, the further the item's difficulty from the examinee's ability level the more information about aberrance provided when it is responded to aberrantly. As item difficulties approach the examinee's ability level, the response probabilities approach .50 and aberrant responses are difficult to distinguish from normal ones. For both indices and both types of aberrance, the effect of Diff on detection rates was

Table 16

Summary of Effects of Four Independent Variables at Three False Positive Rates

	Lz						ECIZ4					
	HIGH ABERRANCE			LOW ABERRANCE			HIGH ABERRANCE			LOW ABERRANCE		
	.01	.05	.10	.01	.05	.10	.01	.05	.10	.01	.05	.10
MAIN EFFECTS												
SAMPSIZ												
High												
Low												
DIFF												
High	B	B	B	B	B	B	B	B	B	B	B	B
Low												
DISC												
High	B	B	B	B	B	B						
Low												
MODEL												
2PLM	B	B	B	B	B		B	B	B	B	B	B
3PLM												
SIMPLE MAIN EFFECTS												
SAMPSIZ												
High												
Low							B for 3PLM	B for 3PLM	B for 3PLM			
DISC												
High										B for 3PLM		
Low												B for 3PLM
ENHANCEMENTS												
High Disc /Low Diff	EN			EN								
2PLM/High Diff	EN	EN	EN				EN	EN	EN	EN	EN	
2PLM/Low Diff												EN

B = Better Detection Rate

EN = Enhanced Detection Rate

greater than for any other independent variable. Logically, the magnitude of this effect is dependent on the choice of the values for the Diff variable. The high condition represented a variance of the b parameter ranging from -2.5 to 2.5, uniformly distributed and lov Diff represented a variance ranging for -1.5 to 1.5. The high range was similar to the range recommended for a general achievement test by Hambleton and Swaminathan (1985). The contrasting low range would be representative of some computerized adaptive tests.

The 2PLM was more effective than the 3PLM for both indices for both high and low aberrance. It appears that the 2PLM is generally more effective than the 3PLM. It was somewhat surprising that the 2PLM was more effective than the 3PLM. Even though Noonan (1989) had obtained the same result, he had studied the effectiveness of the indices for the 2PLM using data for which no guessing had been introduced. In this study the data were generated to fit the 3PLM which includes a guessing parameter. Thus the 3PLM could be expected to fit the data better than the 2PLM.

One possible reason that the 2PLM was more effective than the 3PLM is that for the 2PLM all unexpected responses, whether naturally occurring or as a result of the simulation, were considered aberrant. For the 3PLM, a c parameter which accounts for random guessing is estimated. The magnitude of the effect for the Model variable was quite considerable for high aberrance for both indices, but it was almost negligible for Lz for low aberrance and very low for ECIZ4 for low aberrance (even though detection rates were higher for low aberrance). It may be that when used with the 2PLM, ECIZ4 is better than Lz at detecting random responses. Tables 4 and 5 show that there is little difference in the cutoff values for Lz for the two levels of the Model variable, while for ECIZ4 the cutoff values of the 2PLM are higher than for the 3PLM. For the 2PLM any responses that are not expected are treated as guessing (i.e. as aberrance) and this may cause the rise in the cutoff values for ECIZ4. For high aberrance, where both indices produced better detection rates, the increase in detection rates were higher for ECIZ4, even though the cutoff values for ECIZ4 were higher for the 2PLM than for the 3PLM (Tables 9

and 12). For low aberrance, where the response vectors of simulees of relatively high ability were chosen for modification, there is little guessing and the c parameter of most items is likely to have played little role. This may be the reason that the 2PLM did not have a large impact on detection rates for the low aberrance.

A second possible reason that the 2PLM was more effective than the 3PLM, is that the 2PLM may be giving more accurate results, even though the assumption of no guessing was violated. Accurate estimation of item parameters requires larger numbers of examinees for the 3PLM than it does for the 2PLM. For the 3PLM, item parameters may not be very accurately estimated (particularly the c parameter) even with a sample size of 2500.

If the reason that the 3PLM was less effective is because the item parameters were not estimated accurately enough, there are three conceivable ways of improving the performance of the indices when using the 3PLM. One way is to increase the sample size used to estimate the item parameters. A second way is to increase the number of items in the test. Hulin, et al. (1982) found that in obtaining accurate estimates of item parameters a tradeoff occurred between test length and sample size. A third way is to use an alternative procedure that might be more accurate than LOGIST for estimating the examinee ability and item parameters. However, for conditions that resemble those used in this study, it appears that the 2PLM would be better for high aberrance and although it would not have much effect for low aberrance, the 2PLM would still be more effective.

In general, Sampsiz had little effect on detection rates. The 3PLM requires higher Sampsiz than the 2PLM to obtain accurate parameter estimates. If more accurate parameter estimates lead to better detection, the combination of high Sampsiz and the 3PLM should enhance detection rates.

It is not very clear why Sampsiz had no effect on detection rates. In the only other research to examine the effect of Sampsiz on detection rates, Drasgow (1982) also found no effect. Drasgow suggested that the parameter estimates may be less accurate for lower Sampsiz, but since the parameters are

not affected consistently, they may be randomly less accurate. The errors cancel themselves out and thus have little effect on the effectiveness of appropriateness indices. If the parameter estimates were randomly less accurate for lower Sampsiz, the variance of the detection rates over replications would be higher. However, there was little difference in the standard deviations of the detection rates over replications for the two levels of Sampsiz. More research is necessary to provide an adequate explanation for this result.

For Lz, high Disc was more effective than low Disc for both high and low aberrance, but not for ECIZ4. This suggests that for Lz the higher the discrimination parameter for the items in a test, the higher the detection rate of aberrance. This difference is difficult to explain, but it may be caused by the different formulation of the two indices. ECIZ4 reflects the extent to which an examinee's response vector relates to a theoretical person response curve at a fixed level of theta, while Lz is related to the responses of the entire sample.

While the effect of the level of Disc (the a item parameters within the test) was significant for Lz, the magnitude of the difference in detection rates produced by Disc was only moderate. Hambleton & Swaminathan (1985) have recommended item discrimination parameters for a general achievement test in the range of .40 to 1.50. In this study the item parameters ranged from .8 to 1.4 for high Disc and from .5 to 1.1 for low Disc. These two levels of Disc have a mean difference of .3. It appears that with this magnitude of difference in Disc there was some practical difference in the detection rates produced by Lz, but none was evidenced for ECIZ4. It is possible that a greater difference in Disc would have produced higher differences in the magnitude of the detection rates.

The ANOVA results also showed that the combination of the 2PLM and high Diff produced enhanced detection rates for high aberrance for both indices. As noted above, for both indices, the 2PLM was more effective than the 3PLM and high Diff was more effective than low Diff. These results imply that for

either index, for high aberrance the combined use of the 2PLM with a test that has a broad range of item difficulty parameters enhances the detection of aberrance.

One possible explanation for this interaction is that high Diff increases the number of very easy or very difficult items in the test so that, on the whole, the c parameter is less well estimated. The accurate estimation of the c parameter requires a large number of simulees with ability levels that are similar to the difficulty levels of the items. However, in this study there were relatively few simulees at the extremes of ability because a standard normal distribution had been used to generate examinee ability. This explanation is consistent with the result which showed that the combination of the 2PLM and high Diff did not enhance detection rates for low aberrance. The low aberrance was simulated for examinees with relatively high abilities, while the high aberrance was simulated for examinees with relatively low abilities. Examinees with high ability are less likely to guess, therefore the c parameter plays less of a role for these examinees.

Detection Rates using Cutoff Values from the Standard Normal Distribution

Detection rates for the 16 combinations of Diff, Disc, Model, and Sampsiz were calculated for both indices for high and low aberrance, using cutoff values from the non-aberrant sample and using cutoff values from the standard normal distribution. The differences between the detection rates using each of the two cutoff values were then calculated.

For high aberrance, the absolute difference between detection rates using cutoff values from the non-aberrant sample and from a standard normal distribution are shown in Table 17 for both indices. For Lz, over the 16 combinations of the independent variables, at the .01, .05, and .10 false positive rates respectively, the maximum difference was 6%, 7%, and 7% and the mean absolute difference over combinations was 3%, 2%, and 2%.

Table 17
Absolute Difference Between Detection Rates using Cutoff Values from the Nonaberrant Sample and a Standard Normal Distribution for High Aberrance at Three False Positive Rates

CONDITION				Lz			ECIZ4		
Sampsiz	Diff	Disc	Model	.01	.05	.10	.01	.05	.10
1000	low	low	2PLM	0.041	0.066	0.071	0.040	0.026	0.017
1000	low	low	3PLM	0.009	0.025	0.027	0.019	0.013	0.009
1000	low	high	2PLM	0.014	0.030	0.035	0.028	0.008	0.005
1000	low	high	3PLM	0.008	0.008	0.013	0.007	0.005	0.006
1000	high	low	2PLM	0.032	0.004	0.001	0.039	0.010	0.004
1000	high	low	3PLM	0.032	0.008	0.001	0.031	0.008	0.003
1000	high	high	2PLM	0.058	0.016	0.004	0.062	0.016	0.004
1000	high	high	3PLM	0.035	0.011	0.003	0.020	0.003	0.008
2500	low	low	2PLM	0.046	0.065	0.069	0.022	0.010	0.005
2500	low	low	3PLM	0.017	0.032	0.034	0.004	0.006	0.007
2500	low	high	2PLM	0.019	0.039	0.043	0.013	0.002	0.004
2500	low	high	3PLM	0.002	0.012	0.016	0.008	0.021	0.024
2500	high	low	2PLM	0.031	0.002	0.003	0.036	0.008	0.002
2500	high	low	3PLM	0.029	0.006	0.002	0.017	0.003	0.009
2500	high	high	2PLM	0.054	0.013	0.002	0.047	0.011	0.002
2500	high	high	3PLM	0.034	0.010	0.001	0.016	0.009	0.016
Mean absolute difference				0.029	0.022	0.020	0.026	0.010	0.008
MARGINALS									
Sampsiz	1000			0.01	0.01	0.02	0.03	0.00	0.00
	2500			0.01	0.01	0.02	0.02	0.00	0.01
Diff	low			0.01	0.03	0.04	0.02	0.00	0.01
	high			0.04	0.01	0.00	0.03	0.01	0.00
Disc	low			0.00	0.02	0.03	0.03	0.01	0.00
	high			0.02	0.01	0.01	0.02	0.00	0.01
Model	2PLM			0.01	0.02	0.03	0.04	0.01	0.00
	3PLM			0.01	0.01	0.00	0.02	0.00	0.01

Overall the differences were lower for ECIZ4 for high aberrance. Over the 16 combinations of the independent variables, at the the .01, .05, and .10 false positive rates respectively, the maximum difference was 6%, 3%, and 2% and the mean absolute difference was 3%, 1%, and 1%.

For low aberrance, the absolute difference between detection rates using cutoff values from the non-aberrant sample and from a standard normal distribution are shown in Table 18 for both indices. For Lz, over the 16 combinations of the independent variables, at the the .01, .05, and .10 false positive rates respectively, the maximum difference was 4%, 6%, and 6% and the mean absolute difference was 2%, 2%, and 2%.

As for high aberrance, the differences for ECIZ4 tended to be lower than the differences for Lz. Over the 16 combinations of the independent variables, at the the .01, .05, and .10 false positive rates respectively, the maximum difference was 5%, 2%, and 4% and the mean absolute difference was 2%, 1%, and 1%.

Discussion

In the first section of this chapter the cutoff values obtained by the simulation of a non-aberrant sample were found to be significantly different from those of the standard normal distribution. These differences had been postulated by Molenaar and Hoijtink (1990) and Smith (1982). However, the question of whether these differences would have any practical significance on detection rates for aberrant response patterns has never been researched.

The results of this study show that the use of cutoff values from a standard normal distribution versus the use of cutoff values obtained by the simulation of a non-aberrant sample does not produce a very great difference in detection rates for either index. However, the differences are slightly greater for Lz than for ECIZ4. In the part of this chapter dealing with cutoff values it was observed that the cutoff values that were set using a simulation of non-aberrant response vectors varied considerably over replications and conditions. However, it was also observed that the

Table 18
Absolute Difference Between Detection Rates using Cutoff Values from the Nonaberrant Sample and a Standard Normal Distribution for Low Aberrance at Three False Positive Rates

CONDITION				Lz			ECIZ4		
Sampsiz	Diff	Disc	Model	.01	.05	.10	.01	.05	.10
1000	low	low	2PLM	0.040	0.056	0.055	0.027	0.021	0.038
1000	low	low	3PLM	0.007	0.023	0.028	0.016	0.008	0.007
1000	low	high	2PLM	0.010	0.019	0.033	0.015	0.004	0.004
1000	low	high	3PLM	0.005	0.009	0.022	0.004	0.003	0.017
1000	high	low	2PLM	0.017	0.002	0.001	0.045	0.005	0.001
1000	high	low	3PLM	0.017	0.006	0.001	0.019	0.003	0.017
1000	high	high	2PLM	0.023	0.012	0.001	0.023	0.001	0.000
1000	high	high	3PLM	0.027	0.006	0.000	0.009	0.001	0.004
2500	low	low	2PLM	0.042	0.055	0.053	0.014	0.010	0.011
2500	low	low	3PLM	0.018	0.031	0.036	0.004	0.004	0.010
2500	low	high	2PLM	0.016	0.038	0.035	0.007	0.001	0.009
2500	low	high	3PLM	0.002	0.009	0.030	0.004	0.010	0.042
2500	high	low	2PLM	0.013	0.001	0.002	0.033	0.004	0.001
2500	high	low	3PLM	0.006	0.003	0.003	0.011	0.005	0.018
2500	high	high	2PLM	0.018	0.008	0.003	0.022	0.006	0.001
2500	high	high	3PLM	0.015	0.001	0.001	0.008	0.002	0.024
Mean absolute difference				0.017	0.017	0.019	0.016	0.006	0.013
MARGINALS									
Sampsiz	1000			0.01	0.01	0.02	0.02	0.00	0.00
	2500			0.00	0.01	0.02	0.01	0.00	0.02
Diff	low			0.02	0.03	0.03	0.01	0.00	0.01
	high			0.01	0.00	0.00	0.02	0.00	0.01
Disc	low			0.00	0.02	0.02	0.03	0.00	0.00
	high			0.01	0.01	0.01	0.01	0.00	0.02
Model	2PLM			0.00	0.01	0.03	0.02	0.01	0.00
	3PLM			0.01	0.01	0.01	0.01	0.00	0.02

deviations in cutoff values over conditions is not likely to have a great effect on detection rates. Also, in practice most researchers are likely to do only one replication and use the cutoff value. Under these circumstances, the normal curve cutoff values would almost always fall within one standard deviation of any single obtained value (see Tables 4 and 5). It appears, therefore, that there may be less chance of error if the standard normal distribution were used to set cutoff values.

Summary

In this chapter mean cutoff values at the .01, .05, and .10 false positive rates for Lz and ECIZ4 were examined. In addition, the effects of the Diff, Model, Sampsiz, and Disc independent variables on detection rates and the effects of assuming a standard normal distribution for setting cutoff values were investigated.

The mean cutoff values for Lz were unstable over conditions and replications. In general the instability would not have had a great impact on detection rates. The cutoff values were significantly different from those of a standard normal distribution, but the t - tests used for the analysis had a very large sample size so that very small differences resulted in statistical significance. Each of the four independent variables had an effect on the cutoff values for Lz; certain of these effects were complex and difficult to characterize.

The overall detection rates for Lz were quite high, but they were better for low aberrance than for high aberrance. The detection rates were affected by Diff, Model, and Disc, but not by Sampsiz. High Diff resulted in better detection rates than low Diff and the magnitude of the difference was quite large. The 2PLM produced significantly better detection rates than the 3PLM. However, while the magnitude of the difference was quite large for high aberrance it was not very great for low aberrance. High Disc produced significantly better detection rates than low Disc, but the magnitude of the

difference was not very large. The combination of the 2PLM and high Diff enhanced detection rates for high aberrance.

The detection rates obtained by using a non-aberrant sample for setting cutoff values were compared to those obtained by assuming a standard normal distribution. Little difference was found in these detection rates.

As for Lz, the mean cutoff values for ECIZ4 were unstable over conditions and replications. However, they were more stable than for Lz and, in general, the instability would not have a great impact on detection rates. Over the entire sample the cutoff values for ECIZ4 were significantly different from those of a standard normal distribution at the .01 and .05 false positive rates, but not at the .10 false positive rate. However, the t - tests used for the analysis had a very large sample size so that very small difference resulted in statistical significance. As for Lz, each of the four independent variables had an effect on the cutoff values for ECIZ4; these effects were complex and difficult to characterize and they were different that the effects for Lz.

The overall detection rates for ECIZ4 were generally higher than for Lz. For only one condition, high aberrance at the .01 false positive rate, was Lz more effective than ECIZ4. As for Lz, detection rates for ECIZ4 were better for low aberrance than for high aberrance.

For ECIZ4, the Diff and Model variables had the same effects on the detection rates as for Lz; as for Lz, Sampsiz had no effect on detection rates. However, while for Lz high Disc produced better detection rates than low Disc, there were no significant effects for ECIZ4. As for Lz, the combination of the 2PLM and high Diff enhanced detection rates for high aberrance.

There was not a large difference in detection rates obtained by assuming a standard normal distribution for setting cutoff values compared to the detection rates obtained by using a non-aberrant sample to set cutoff values. The difference was less for ECIZ4 than for Lz.

CHAPTER VI

CONCLUSIONS

In this chapter general conclusions based on the research are discussed, limitations of the research are presented and suggestions for future research are made.

Most of the simulation research on appropriateness indices has not involved any replications. The first conclusion reached from this research is the advisability of replicating conditions to avoid chance results. The research showed that cutoff values and detection rates can vary considerably over replications.

The second conclusion is that based on the present evidence, if a choice of one index must be made for practical applications, the ECIZ4 should be used. There are a number of reasons for this conclusion. The cutoff values for ECIZ4 were more stable over replications than for Lz. If a practitioner wishes to generate a sample of non-aberrant response vectors for setting cutoff scores, more stable results would be obtained with ECIZ4 for a given number of replications. Or, if the practitioner wishes to use cutoff values from a standard normal distribution, the difference in detection rates between those that would be obtained from setting cutoff values by generating a sample of non-aberrant response vectors would be less for ECIZ4 than it would be for Lz. Finally, ECIZ4 generally had higher detection rates than Lz (although within combinations of conditions, detection rates were slightly more variable over replications for ECIZ4).

The third conclusion is that there is not a great difference in detection rates for Lz and ECIZ4 when cutoff scores from a standard normal distribution are used versus those set by simulation and for practical applications it may be best to use the standard normal distribution. This conclusion is based on the results of the study which produced only small differences in the detection rates for the two methods of setting cutoff

values. In addition, previous research has shown that Lz and ECIZ4 both followed a standard normal distribution quite closely. As mentioned above the difference is less for ECIZ4 than for Lz.

The fourth conclusion is that the greater the variance of the b parameter of the test items, the higher the detection rate. This conclusion was suggested by the research of Reise and Due (1991), but the methodology used was different from that of the present study and only the Lz index was used. In the present study, the variance of the b parameter affected detection rates for both indices and for high and low aberrance. A strong rationale exists for this conclusion.

In addition to the above conclusions the following are tentative conclusions that have limited generalizability because of the limitations of the study:

The fifth conclusion is that the appropriateness indices are likely to be more useful for detecting low aberrance than for high aberrance. In general low aberrance was better detected than high aberrance, although there were some combinations of conditions for which high aberrance was better detected.

The sixth conclusion is that the use of sample sizes of 1000 versus sample sizes of 2500 for the estimation of item parameters has no effect on detection rates for the Lz and ECIZ4 indices. This conclusion is based on the results of the ANOVAS which produced no significant main effect or interactions for this factor. It is noted that significant effects might have been produced with smaller sample sizes.

The seventh conclusion is that for both indices the 2PLM is more effective than the 3PLM for the detection of both high and low aberrance. This conclusion is based on the results of the ANOVAS which generally indicated that the 2PLM was significantly more effective than the 3PLM at the .01, .05, and .10 false positive rates. However, it is noted that the magnitude of the effect was quite low for low aberrance.

The eighth conclusion is that the combination of the 2PLM and high variance of the b parameter enhances detection rates producing an ordinal interaction of high aberrance for both indices. This conclusion is based on the results of the ANOVAS which were significant at all three false positive rates.

The ninth conclusion is that higher item discrimination parameters in a test produce higher detection rates for Lz, but not for ECIZ4. This conclusion is based on the results of the ANOVAS which were significant at all three false positive rates. However, it is noted that the magnitude of the effect was quite low.

Limitations of the Study

The generalizability of the results of this study are subject to a number of limitations. First, it must be recognized that the simulation did not fully represent all of the conditions that would be present in a real setting. In particular, the fit of the model to the data was not examined. In addition, the item parameters were computer generated from specific ranges and may not represent those of a real test. Second, only one test length was used and only one level of aberrance. Third, all the levels of the independent variables were dichotomous. For some of the variables, such as sample size, a systematic relationship may exist for detection rates which might have been apparent if more levels were used. In particular, it would have been useful to have included a sample size below 1000 (perhaps in the 400 to 500 range). In a real test situation researchers may be required to "choose their poison" - a misfitting model (i.e., the one or two parameter model) with reasonable parameter estimates and a better fitting model with relatively poor parameter estimates. Fourth, the study was limited to the 2PL and 3PL models. Results with the 1PL might have revealed some systematic results and have aided interpretation.

Suggestions for Future Research

The complex design of this study, involving four fully crossed factors, increases the generalizabilty of the results. However, as pointed out by Cohen and Cohen (1983), complex designs, such as this, often lead to results that are difficult to interpret. It is suggested that a number of questions be examined for a fuller understanding of their causes and to permit greater generalizability.

Low aberrance was better detected than high aberrance. This result was most likely caused because little guessing occurs for high ability levels and thus a departure from expected values is more notable than it is for high aberrance. However, it appears that the item characteristics may play an important role in determining the effectiveness of the indices for high and low aberrance and the effect of these characteristics should be explored more fully.

Sample size had no effect on detection rates and it has been suggested that while the parameter estimates may be less accurate for lower sample size, the parameters are not affected consistently so that errors in the parameter estimates cancel themselves out and have little effect on the effectiveness of the indices. However, there was some evidence in this study that these parameter estimates were not randomly less accurate. More extensive research is necessary to clarify this issue. In particular, it would be useful to examine this issue using smaller sample sizes than were used in this study.

The 2PLM was more effective than the 3PLM. One possible reason that the 2PLM was more effective than the 3PLM is that for the 2PLM all unexpected responses, whether naturally occurring or as a result of the simulation, were considered aberrant. It appears that ECIZ4 was better at detecting this random aberrance than Lz. To help verify this rationale, it would be interesting to compare the two indices when the aberrance was not random. Another reason that the 2PLM was more effective than the 3PLM may be that the item parameters were not estimated accurately with the 3PLM. The estimation of the item parameters for the 3PLM may be improved by increasing the sample

size used to estimate item parameters, increasing the number of test items, and/or using a different procedure for estimating examinee ability and item parameters. Since in some practical applications it may be desirable to use the 3PLM, the effects of these variables on the detection rates of the indices should be examined.

The combination of the 2PLM and high range in the distribution for the b parameter enhanced detection rates for high aberrance, but not for low aberrance. In this study, only half of the ability range was used and it is possible that this result may partly have been caused by the range of ability used to generate aberrant responses. Low aberrance may not have been affected as much because simulees with high ability are less likely to guess, therefore the c parameter plays less of a role for these examinees.

Additional suggestions for research follow from the limitations of this study. It is suggested that some of the questions posed in this study be examined using real data. To permit greater generalizability, it would also be useful to examine the effects of the independent variables that were used in this study with different test lengths and for different levels of aberrance. To permit clear trends to appear, it would be useful to examine the effects of some of the independent variables, such as sample size and Disc, using a number of levels. In addition, it would be interesting to study the effect of the independent variables when using the 1PLM.

Finally, it is noted that no researcher has systematically studied the extent to which Lz and ECIZ4 detect the same aberrant response vectors and how different conditions affect this detection. This research is suggested for a fuller understanding of how the indices perform and under which circumstances each is likely to be most effective. In addition, no researcher has examined the relationship of test information to the performance of the indices under different conditions. Test information would assist in the interpretation of the results. For example, the test information function could help provide a clearer understanding of the reasons for the effectiveness of the indices under different conditions of the independent variables.

REFERENCES

Anderson, M. (1990). Scoreup: A Computerized Item Analysis and Examinee Management System. Reference manual available from Martin W. Anderson, 49 Meadow Farms Road, West Harford, CT 06107.

Birenbaum, M., & Tatsuoka, K. K. (1982). On the dimensionality of achievement test data. Journal of Educational Measurement, 19, 259-266.

Carlson, J. (1985). IBM version of DATAGEN. University of Ottawa.

Chatman, S. P. & Nelson (1984). The influence of guessing on measures of response aberrance when using the rasch model. Internal report. Texas A&M University. Measurement and Research Services.

Cohen, J., & Cohen, P. (1983). Applied Multiple Regression/Correlation Analyses for the Behavioral Sciences. Hillsdale, N.J.: Lawrence Erlbaum Ass.

Cronbach, L. T. (1946). Response sets and test validity. Educational and Psychological Measurement, 6, 475-494.

Donlon, T. F., & Fisher, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. Educational and Psychological Measurement, 28, 105-113.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6, 297-308.

Drasgow, F. (1985). A computer program to compute three appropriateness indices.

Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. Journal of Applied Psychology, 72(1), 10-18.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. Applied Psychological Measurement, 10, 59-67.

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

Drasgow, F., Levine, M.V., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. (1987). Modeling incorrect responses to multiple-choice items with multilinear formula score theory.ERIC document (ED 654 876).

Fischer, F. E. (1970). Some properties of the personal biserial index. Journal of Educational Measurement, 7, 275-277.

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

Hambleton, R.K., & Rovinelli, R.J. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 17, 73-74.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: principles and applications. Boston: Kluwer-Nijhoff.

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.

Harnish, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. Hambleton (Ed.), Application of item response theory. Vancouver: ERIBC.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin, 105(2), 320-308.

Hulin, C. L., Drasgow, F., & Parson, C. K. (1983). Item response theory: Application to Psychological Measurement. Homewood, IL: Dow Jones-Irwin.

Hulin, C. L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two and three parameter logistic item characteristic curves: A Monte Carlo study. Applied Psychological Measurement, 6, 249-260.

Keppel, G .(1991). Design and Analysis: A Researcher's Handbook. New Jersey: Prentice Hall.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), New Horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Levine, M. V., & Drasgow, F. (1984). Performance envelopes and optimal appropriateness measurement (Report 84-5). Champaign, ILL University of Illinois, Department of educational Psychology, Model-Based Measurement Laboratory. (ERIC Document No. ED 263 126)

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. Psychometrica, 53, 161-176.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics, 4, 269-290.

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

Lumsden, J. (1980). Variations on a theme by Thurstone. Applied Psychological Measurement, 4, 1-7.

Molenaar, I.W. & Hoijtink, H. (1990). The many Null distributions of person fit indices. Psychometrika, 49(1), 75-106.

Noonan, B. (1989). The effect of test length, IRT model, type of aberrance, and level of aberrance on the distribution and effectiveness of three appropriateness indices. Unpublished doctoral dissertation. University of Ottawa. Ottawa, Ontario, Canada.

Oltman, P. K. (1985). Background characteristics of examinees showing unusual test behavior on the Graduate Record Examination. (ETS Research Report 85-39). Princeton, N. J.

Reise, S.P., & Due, A.M. (1991). The influence of test characteristics on the detection of aberranct response patterns. Applied Psychological Measurement, 15, 217-226.

Rudner, L. M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

Rosenthal, R., & Rosnow, R. (1991). Essentials of Behavioral Research: Methods and Data Analysis. Toronto: McGraw-Hill, Inc.

Sato, T. (1975). The construction and interpretation of S-P tables. Tokyo: Meiji Tosho. (in Japanese)

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

Smith, R. M. (1982). Detecting measurement disturbances with the Rasch model. Unpublished Ph. D. Dissertation, University of Chicago.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. Educational and Psychological Measurement, 45, 433-444.

Smith, R. M. (1986). Person fit in the rasch model. Educational and Psychological Measurement, 46, 359-372.

Tatsuoka, M. M. (1979). Recent psychometric developments in Japan: Engineers tackle educational measurement problems.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrica, 49(1), 95-110.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7, 81-96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7, 215-231.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 20, 221-230.

Tatsuoka, K. K., & Tatsuoka, M. M. (1985). Bug distribution and pattern classification. ERIC document (ED 263 147).

Trabin, T. E., & Weiss, D. J. (1979). The person response curve: The fit of individuals to item characteristic curve models Research Report 79-7. Minneapolis: University of Minnesota, Psychometric Methods Program.

Van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y. H. Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swets and Seitlinger, B. V.

Weiss, D.J. (1973). The stratified adaptive computerized ability test (Research report Report 73-3). Minneapolis: University of Minnesota, department of Psychology, Psychometric Methods Program, (NTIS No. AD 768376)

Wingersky, M.S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R.K. Hambleton (Ed.), Applications of item response theory (pp. 45-56). Vancouver B.C.: Educational Research Institute of British Coluumbia.

Wingersky, R.M., Barton, M.A., & Lord, F.M. (1982). LOGIST user's guide. Princeton, N.J.: Educational Testing Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch Model. Journal of Educational Measurement, 14, 97-116.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Zimmerman, D. W. & Zumbo, B. D. (1992). Correction for nonindependance of sample observations in ANOVA F tests. Journal of Experimental Education, 60, 367-381.