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.

The Effect of Test Length, IRT Model, Type of Aberrance, and Level of Aberrance on the Distribution and Effectiveness of Three Appropriateness Indices

Brian W. Noonan

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

University of Ottawa

ISBN 0-315-59995-2

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ABSTRACT

A test score which does not measure an examinee accurately may be said to be inappropriate for that examinee. Researchers have suggested that an examinee's response vector, the pattern of correct and incorrect responses, may contribute information not available in the total test score. Analysis of response patterns is proposed as one way to assess the appropriateness of an examinee's test results because an unexpected or unusual pattern may indicate that the individual has been mismeasured.

Recently quantitative measures called appropriateness indices have been developed to detect unusual response patterns. A number of researchers have investigated detection rates of some of the appropriateness indices. It has been found that the standardized versions of the indices based on item response theory may be more accurate and consistent in detecting unusual response patterns than are others such as group dependent indices. Under the conditions which have been investigated, three indices, i) Lz, ii) ECIZ4, and iii) W3, have been found to be among the most effective.

One problem related to appropriateness indices is the effect of test length and IRT model on the characteristics of the distributions of the indices. Studies related to the effect of test length have shown that appropriateness indices tend to be more effective in identifying aberrant response

patterns in longer rather than shorter test lengths. However, in those studies type of test was confounded with test length. Also, only the 3PLM and the Rasch Model have been compared and then only with non-standardized indices. A related question is the extent to which the characteristics of the distributions are stable over samples.

A second important problem is the extent to which the detection rates of the indices are affected by combinations of test length, IRT model, type of aberrance, and level of aberrance. Although researchers have investigated the effects of some of those variables, the effects of combinations of those variables have not been systematically examined.

There were two basic purposes for this study. The first purpose was to investigate the characteristics of the distributions of Lz, ECIZ4, and W3 for non-aberrant response patterns in combinations of test lengths (40 items and 80 items) and IRT model (the 2PLM and the 3PLM). The second purpose was to investigate the effectiveness of the three indices in twenty-four combinations of two test lengths, two IRT models, two types of aberrance, and three levels of aberrance.

In order to investigate the distributions of appropriateness indices in non-aberrant response patterns, data were generated by computer to simulate various measurement conditions. Item parameters were generated within specified ranges to produce similar tests for the two test lengths and two IRT models.

Simulated examinees were generated from the normal (0,1) distribution. Two thousand non-aberrant, response vectors were generated for each of four conditions, test length by IRT model. The three appropriateness indices, Lz, ECIZ4, and W3 were calculated for each examinee. This procedure was replicated fifty times for each of the four combinations of test length and IRT model. Of the three indices, ECIZ4 produced the most stable distributions over replications.

To examine the effect of test length and IRT model on characteristics of the distributions of the indices, the mean, standard deviation, skewness, and kurtosis were computed for each index in each of the combinations of test length and IRT model over fifty replications. There were no significant effects for either test length or IRT model on the means of the three indices. Based on skewness and kurtosis, the distributions of ECIZ4 most closely approximated normality, while the distribution of W3 was least normal. To establish false positive rates, the tails of the distributions of each index were then examined at P_{01}, P_{05}, P_{10}, and P_{25} for each of the four conditions. Of the three indices ECIZ4 seemed least affected and W3 most affected by test length, IRT model, and the interaction of test length and IRT model.

To investigate the effectiveness of the indices, aberrant response patterns were generated for the twenty-four combinations of the four variables (2 test lengths x 2 models x 2

types of aberrance x 3 levels of aberrance). Four thousand simulated examinees were generated for each of the twenty-four combinations and each index was computed for each examinee for each of the twenty-four combinations. The detection rates of the indices were then computed and compared for each index for each of the twenty-four conditions.

Overall, the 80 item test produced somewhat better detection rates than the 40 item test and the 2PLM better rates than the 3PLM. Spuriously low scores tended to produce slightly higher detection rates than spuriously high scores under most conditions. Higher levels of aberrance tended to produce higher detection rates although for some conditions there was little difference between 15% and 30% aberrance. Lz and ECIZ4 tended to produce better detection rates than W3; however, no detection rates seemed to be as high as those reported in previous research.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the guidance and direction which Dr. Marvin Boss provided in the preparation of this thesis. His generous sharing of time and expertise was invaluable. Special appreciation is reserved for the author's wife, Donna, who provided the support and encouragement which made completion of the project possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

CHAPTER 1

REVIEW OF RESEARCH

Valid interpretation of individual examinee test scores is a fundamental concern in educational measurement, especially when test scores serve as the basis for decisions about individuals. Test score interpretations which are used for various forms of candidate selection, certification, or mastery learning may have serious consequences, not only for the examinee but for those making decisions about test scores. Consequently, the exploration of various aspects of individual examinee measurement has become a topic of interest in educational and psychological measurement.

Sometimes there are forms of item-examinee interaction that can result in an individual being "mismeasured", which means the test score cannot be validly interpreted. For example, item bias is said to exist when an item is found to be differentially difficult for a particular sub-group which would result in members of the group being measured inaccurately. A related topic of current interest is the detection and interpretation of unexpected, unusual, or aberrant test response patterns. This is a form of examinee mismeasurement which can be considered as analagous to item bias, except that, rather than one or more items mismeasuring a group, a series of items (i.e. a test)

mismeasures an individual. One of the indicators of mismeasurement is the extent to which an examinee's response pattern can be considered unusual or unexpected.

The purpose of this review is to describe the measures which may be used to detect unusual response patterns, to raise questions related to existing research, to identify problems for further investigation, and from these problems to select a specific problem to be investigated. The first section is a review of some of the basic principles underlying the concept of person measurement. The second section includes an explanation of the techniques and indices which have been developed to measure unusual response patterns of individual examinees. In the third section, research related to appropriateness measurement is reviewed and questions with respect to existing research and problems needing research are raised. The chapter concludes with a problem statement and research questions.

Person Measurement

Although studies of examinee response patterns and other aspects of person measurement are fairly recent, interest has been expressed for some time in the basic concept. Ghiselli (1960, 1963) in discussing test validity suggested that predicting predictability was needed to ensure precision in test score interpretation. It was suggested that test validity could be improved if an index or moderator were used when predicting a criterion measure. This was one of the earliest attempts to

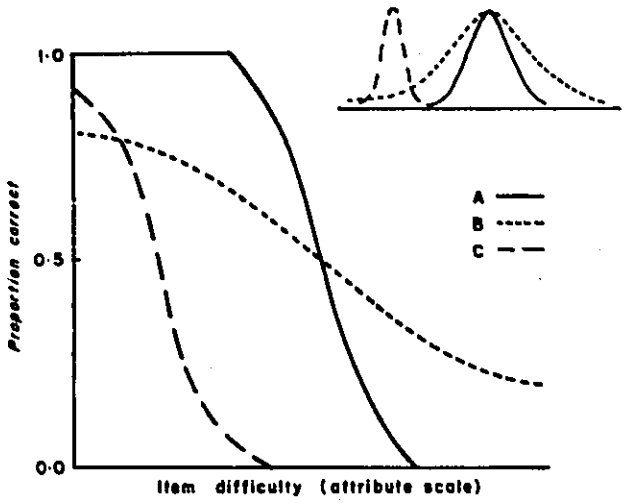
consider the effects of item-examinee mismeasurement and to recognize that individual characteristics may require different interpretation of scores.

Lumsden (1977, 1978) examined theoretical aspects of inconsistent or unusual individual examinee response patterns. Lumsden (1977) proposed that three types of psychological fluctuations, developmental trends, swells, and tremors, could help to explain measurement disturbances which may contribute to an unreliable test score. Developmental trends are described as fluctuations in ability developed over a period of time, such as changes due to learning or cognitive growth. Swells are short term fluctuations resulting perhaps from mood shifts, an environmental influence, or a stimulus such as a test. Tremors are even shorter term fluctuations occurring perhaps from item to item on a test. In relation to test theory, each of these types of fluctuations can be considered intra-individual because they are sources of unreliability within the examinee rather than in the test. Lumsden (1978) proposed that these fluctuations could be measured by analyzing person characteristic curves (PCC) and that such measurement could be considered as a form of personal test reliability.

The PCC is analogous to the item characteristic curve in that it is a plot of proportion passed by an individual plotted against item difficulty. Figure 1 shows the general form of three observed PCCs in which item difficulty is considered as a continuous variable in the way that examinee ability is

Figure 1

Hypothetical Person Characteristic Curve for Three Subjects



Note: From "Tests are perfectly reliable" by J. Lumsden, 1978, British Journal of Mathematical and Statistical Psychology, 31, p. 21. Copyright 1978, The British Psychological Society, London. Reprinted by permission.

considered continuous in an item characteristic curve.

As shown in Figure 1 the slope of an observed PCC can vary from person to person on a test. If persons A and B are of equal ability, one interpretation is that person A gets easy items correct and more difficult items incorrect; thus, the PCC might be considered an expected curve for that ability. Person B, however, gets some easy items incorrect and more difficult items correct and therefore the PCC might be considered unexpected. Also it can be seen that person C, though of lower overall

ability, performs better than person B on easy items and, on those items, might be considered more able. In analyzing each PCC the focus for item-examinee interaction can be shifted from the item (or test) to the examinee.

Trabin and Weiss (1979), in a study of examinee response variability, explored some of the properties of the PCC which they labelled the Person Response Curve (PRC). Using data from a psychology test they developed a technique to compare observed and expected person response curves for 151 examinees. The results of the study indicated that the PRC could be useful in studying the fit of individuals to the three parameter logistic model. The researchers suggested that the PRC might be used to investigate forms of item-examinee interaction.

A different approach to the application of person measurement has been reported by Carroll, Meade, and Johnson (1986). The focus of this research was the use of person measurement to examine the construct validity of a test rather than the response patterns of individual examinees. Because this approach is not directly related to the analysis of unusual response patterns it will not be explored further in this review.

Sato (1975), as reported by Tatsuoka and Tatsuoka (1982), Harnisch and Tatsuoka (1983), and Tatsuoka (1984), is considered to be one of the first to employ a systematic, mathematical approach to the analysis of unusual response patterns. Sato utilized a method to assess individual response variability in classroom test situations. The method is based on the principle

that it is possible to measure the extent to which an examinee's pattern of responses can be considered unusual within a group. This may be achieved by comparing the order of actual patterns of an examinee's responses with that of an expected order of response patterns for the total group.

In summary, examinee response variability can result in examinee response patterns which can be considered unusual or unexpected. In the next section a number of techniques used to measure unusual response patterns are presented and discussed.

Measurement of Unusual Response Patterns

Analysis of unusual response patterns is of interest because such analysis may contribute information not available in the total test score. Furthermore there may be many plausible response patterns for a typical multiple choice test. Harnisch (1983) observed that there are over two hundred possible correct/incorrect response patterns for a score of five on a ten item test. Recently there has been considerable interest in exploring the nature and extent of unusual response patterns in instruments such as multiple choice tests where dichotomous scoring may be used. In addition it is recognized that an examinee's test score may be invalid even if the test has satisfactory measurement properties for the group being measured (Drasgow & Guertler, 1987). A test score which measures an individual inaccurately can not be validly interpreted and may be considered inappropriate. Analysis of response patterns is one

way to assess the appropriateness of an individual test score.

There can be many reasons for unusual response patterns and inappropriate individual test scores. Inappropriate scores may be considered spuriously high in some instances or spuriously low in others. Wright (1977) suggested that guessing, cheating, or coaching may produce spuriously high scores while "sleeping", plodding, or cultural bias may produce spuriously low scores. Hulin, Drasgow, and Parsons (1983) also suggested that unintentional errors on answer sheets and novel or creative interpretations of items are other possible sources of spuriously low scores. Other situational or contextual factors such as atypical curricula, language problems (Rudner, 1983), and test anxiety (Schmitt & Crocker, 1984) may also produce unusual response patterns which may not be validly interpreted.

A number of techniques have been developed to measure unusual response patterns. Each technique produces a quantitative measure in the form of a coefficient or index, which indicates the extent to which an individual examinee may be mismeasured on a particular test. For the purposes of this research these techniques collectively will be referred to as appropriateness indices.

Several different classifications of indices have been presented. Hulin, Drasgow, and Parsons (1983) suggested two categories, heuristic and item response based. Tomsic (1986) identified three types of indices, group-dependent, item response based, and extended caution indices. Harnisch and Tatsuoka

(1983) suggested two categories, item response based and patterns of right and wrong answers. Harnisch (1983) also used two categories, group-dependent and item response based, which will be used in the following two sub-sections because these categories more clearly distinguish the principles which underlie the approaches. First, group-dependent indices which use standard summary statistics from binary response data are reviewed; then IRT indices which utilize a model-based approach to appropriateness measurement are reviewed.

Group-dependent Indices

One of the first measures used to examine response pattern consistency was the personal biserial correlation, PERBIS, proposed by Donlon and Fischer (1968). PERBIS is a biserial correlation of an examinee's response vector with the proportion correct vector of the examinee group. If an examinee's response pattern is consistent with that of the group, this will yield a high positive correlation. Low or negative correlations indicate that the response pattern is not similar to the group determined item difficulties. Fischer (1970) used PERBIS with two samples on the SAT-V but found that it did not improve prediction of college success. More recently, PERBIS has been criticized as an appropriateness measure on the basis that correlations in general are dependent on the total test score (Hulin, Drasgow, & Parsons, 1983). Although PERBIS was not found to be a practical measure it does indicate the general principle that a group dependent index is derived from an analysis of the relationship between an

examinee's response and item difficulties as determined by a specific test group. Harnisch and Linn (1981) also described the personal point-biserial, which is the product-moment correlation between the item scores for an examinee and the item difficulty for each item. A high correlation suggests that the examinees correctly answer items which are easy for the group and incorrectly answer those which are difficult for the group. The limits of the index depend upon the examinee's actual number right, thus making it dependent also on examinee ability.

The most commonly used group-dependent index is the caution index based on Sato's Student - Problem (S-P) Curve (Harnisch & Linn, 1981; Tatsuoka & Linn, 1983). One of the original purposes of the caution index was to assist classroom teachers in analyzing test results by identifying scores which were not typical, consistent, and expected for an examinee. Figure 2 shows a hypothetical S-P curve illustrating the response patterns of 18 examinees to a 5 item test.

In the example, the subjects are arranged in descending order of scores and the items in ascending order of difficulty, left to right. To measure unusual response patterns for each person, vertical lines are drawn to indicate the total correct score and all the lines are joined. For each item a dotted horizontal line is drawn under the item number to represent the number of examinees who answered it correctly. As a result the S curve (solid line) is the step function of the cumulative distribution of total scores Yi and the P curve (dotted line) is

the corresponding function of the correct answers to the items (Tatsuoka & Linn,1983). For example, Figure 2 shows that subjects 3 and 7 display expected response patterns, since easy items are answered correctly and more difficult items incorrectly.

Figure 2

Hypothetical Example of an S-P Table

Examinee i	Item j 1	2	3	4	5	Examinee Total $n_{i.}$
1	1	1	1	1	0	4
2	1	1	1	0	1	4
3	1	1	1	0	0	3
4	1	1	0	1	0	3
5	1	1	0	0	1	3
6	1	0	1	0	1	3
7	1	1	0	0	0	2
8	1	1	0	0	0	2
9	1	0	1	0	0	2
10	1	0	0	1	0	2
11	0	1	1	0	0	2
12	0	1	0	1	0	2
13	1	0	0	0	0	1
14	1	0	0	0	0	1
15	0	1	0	0	0	1
16	0	0	1	0	0	1
17	0	0	0	1	0	1
18	0	0	0	1	0	1
$n_{.j}$	12	10	7	6	3	

Note. From " Analysis of Item Response Patterns; Questionable Test Data and Dissimilar Curriculum Practices " by D.L.Harnisch and R.L. Linn, 1981, Journal of Educational Measurement, 3, p.136. Copyright 1981, American Educational Research Association, Washington, D.C. Reprinted by permission.

Conversely, subjects 6 and 17 incorrectly answered an easy item and correctly answered a more difficult one. The caution index measures the extent to which such response patterns can be

considered unusual or unexpected.

In its general form the caution index compares the actual student and problem curves to those that would occur if there was a perfect S or P curve. When the curves are perfect, they are identical. This would be the case if there was a perfect correlation between examinee scores and item difficulty which would be unlikely in reality. From the example in Figure 2, the index, Ci, can be written as;

$$C_i = \frac{\sum_{j=1}^{n_{i.}} (1 - U_{ij}) n_{.j} - \sum_{j=n_{i.}+1}^{J} U_{ij} n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - n_{i.} \left(\frac{\sum_{j=1}^{J} n_{.j}}{J} \right)}$$

where: i indexes the examinees, 1 ..I.
j indexes the items, 1..J.
U_{ij} = observed examinee response.
$n_{i.}$ = total correct for examinee i.
$n_{.j}$ = total correct for item j.

The index is a positive number with a large value indicating that an examinee's response pattern deviates from what might be expected and therefore should be interpreted with caution.

Harnisch and Linn (1981) proposed a modified caution index (MCI) which is the original caution index scaled between 0 and 1 to eliminate extreme values.

The MCI can also be written as the ratio of two covariances as follows:

$$MCI = 1 - \frac{\text{cov}\ (\underline{Y}_i\ \underline{Y}_\cdot)}{\text{cov.}(\underline{X}_i\ \underline{Y}_\cdot)}$$

where $\underline{Y}_i$ is the observed response vector for an examinee
$\underline{Y}_\cdot$ is the vector of column total
$\underline{X}_i$ is the Guttman-scaled vector of expected responses.

The numerator is the covariance of the examinee's response pattern and the actual group pattern; the denominator is the covariance of the examinee's expected response pattern and the actual group pattern. Buxie (1986) reported the distribution of the MCI to be approximately normal under various conditions of test length, item difficulty, and simulated guessing.

A number of indices which are similar to the MCI have also been proposed. Kane and Brennan (1980) and Brennan and Kane (1977) developed a dependability index for use with domain referenced tests. This index is a measure which compares the actual agreement of an individual's scores on two randomly parallel tests with the maximum agreement possible. It has been shown to be linearly related to the MCI (Harnisch & Linn, 1981). Because of the similarity between this group-dependent index and the MCI it will not be discussed further here.

Two other group-dependent measures, the Norm Conformity Index (NCI) and the Individual Consistency Index (ICI), have been reported by Tatsuoka and Tatsuoka (1982a). The NCI enables comparison of an individual examinee's response pattern with that of a total group. The ICI is a repeated measures index used to determine whether an examinee's response pattern varies from one

parallel test to another. These indices measure consistency of individual responses to a test or series of tests. For test items which are ordered from easy to difficult, the NCI can be calculated as follows:

$$NCI = 2\ U_{ia}/U_1$$

In this formula U_{ia} is the number of 0's (incorrect responses) to the right of each 1 (correct response) in the examinees' response vector and U_1 is the product of the number of 1's and 0's in the response vector. The index can range from -1 to 1 where 1 is perfect agreement with the group determined item difficulty order. In this way the NCI compares responses to the item difficulty order for a reference group.

The ICI is calculated similarly to the NCI except the item order is determined by an examinee's scores on three or four parallel tests. An index which approaches 1, indicates that an examinee is responding consistently. For purposes of test score interpretation, using an ICI with a total score can help assess the extent to which a student has mastered a particular concept. Tatsuoka and Tatsuoka (1982a) used the ICI with 127 eighth grade math students and found the index to be useful in identifying students who repeatedly mis-applied math rules. Buxie (1986) investigated the distribution of ICI with respect to the influence of test length and simulated guessing. It was found that the index produced high kurtosis values which suggested that the distribution of the index was non-normal.

Van der Flier's U has also been proposed as a group-

dependent index (Harnisch & Linn, 1981). The index U is calculated from an item-examinee order similar to the S-P table by adding the number of 1's (correct responses) to the right of each 0. The index has a range of values between zero and one.

In summary, group-dependent indices are useful measures in specific situations such as with teacher-made tests or with the assessment of curriculum or instructional effects. Of these indices, the MCI, NCI, and ICI have been found to be useful in situations where samples may be relatively small.

Item Response Theory Based Indices

Item response theory (IRT) provides a theoretical basis for the second approach to the identification of unusual examinee response patterns. The shape of the usual item characteristic curve is a logistic curve which follows the familiar S shape and can be written:

$$Pi(\theta) = c + \frac{(1 - c)\, e^{1.7a(\theta - b)}}{1 + e^{1.7a(\theta - b)}}$$

where "a", "b", and "c" are the discrimination, difficulty, and pseudo-chance parameters for item i. Pi(θ) is the probability of a correct response given theta and 1.7 is the scaling factor constant. Two and one parameter IRT models can be estimated by setting "c" equal to zero in each model and setting "a" equal to 1 in the one parameter model. A strength of IRT is that ability estimates can be made independently of the examinee group and that once item parameters are known an expected test score based

on probability can be assigned to each ability level. In the next section a review of three types of appropriateness indices based on the principles of item response theory are presented and discussed.

Maximum likelihood function estimates. Levine and Rubin (1979) investigated the relationship between IRT and unusual test response patterns. Using maximum likelihood estimation, these researchers proposed four indices to measure response atypicality (Drasgow, 1982; Hulin, Drasgow, & Parsons, 1983; Levine & Drasgow,1984; Levine & Rubin, 1979). The general maximum likelihood estimate index Lo can be written as:

$$Lo = \log \max \text{Prob}\ (U \mid \theta)$$

where U is a vector of item responses and θ the maximum likelihood estimate of ability. Lo and the related indices are based on the principle that if an individual's pattern of responses does not contribute to maximizing the likelihood function, the examinee's score is not appropriate for an individual representing the group being measured (Levine & Rubin, 1979). Levine and Drasgow (1982, 1984) have conducted a number of studies on the derivation and possible application of Lo and the related appropriateness measures.

An underlying assumption for Lo is that ability does not fluctuate from item to item. For an unusual response pattern, the value of Lo would typically be a small negative number (i.e. -30 or less). The value of Lo can be affected by responses omitted or not reached; therefore a second index, the geometric

mean Lm, was suggested. Lm is based only on the items to which an examinee responded. A third index, Lg, sometimes written Ln, was suggested to account for examinee ability fluctuations from item to item (Levine & Drasgow, 1982). Lg assumes a Gaussian model with an ability estimate (θ) and variance (σ^2) for each item. The fourth index is a likelihood ratio test, LR, which can be written either as LR = Lg/Lo or as LR = Lg - Lo. A large LR would indicate a relatively unimportant amount of variation in the ability estimates (Levine & Drasgow,1982). A more complete discussion of the assumptions and derivations of the Lo type indices is contained in Hulin, Drasgow, and Parsons (1983).

To improve the usefulness of the indices, Levine, Drasgow, and Williams (1985) introduced a standardized form of Lo which can be written either as Lz or Z3. Lz is based on the conditional mean and standard deviation of Lo as follows: $Lz = Lo - \underline{E}(\theta) / \underline{\sigma}(\theta)$ where $\underline{E}$ and $\underline{\sigma}$ are the vectors of the conditional mean and standard deviation. A second standardized index, Zh, with properties similar to Lz, has been suggested for the analysis of polychotomous responses and for examinees with high omit rates.

An important reason for developing Lz was to reduce the dependence of the index on examinee ability. Drasgow, Levine, and Williams (1985) found that the standardized indices, Lz and Zh, were also slightly related to ability but not so as to be important in interpreting the indices. They also found that the distributions of the Lz and Zh indices were approximately normal.

Other non-standardized IRT based indices have been utilized. Drasgow, Levine, and McLaughlin (1987) used two indices, the jackknife and the observed/expected curvature which are based on the assumption that the likelihood function will have a flattened curve near its maximum indicating a poor fit. Also, Levine and Rubin (1979) have used the square root of the maximum likelihood of the ability variance as an appropriateness index.

More recently Drasgow, Levine, and McLaughlin (1987) and Drasgow and Levine (1986) have proposed the concept of the optimal index which is considered to be a theoretical index such that no other could achieve higher detection rates and to which others might be compared. Using the Neyman Pearson Lemma they proposed that the likelihood ratio is as powerful as any that can be computed. The general form of their optimal index is written $LR = P_{aberrant}(\underline{u}) / P_{normal}(\underline{u})$ where $\underline{u}$ is the response pattern. The researchers emphasized that LR is of value as a research tool only because to compute the index the form of aberrance must be fully specified, which is not a realistic condition for practical purposes.

In summary, maximum likelihood function indices are derived from IRT and as such provide a theoretical basis for measuring unusual response patterns. They are based on the assumption that a more able examinee will have a higher probability of passing a more difficult item than will a less able examinee. This enables the calculation of an index for an examinee independent of the

group response pattern. Finally, it has been shown that Lz, a standardized version of Lo, is relatively independent of ability.

Fit statistics. A second general category of IRT based measures is associated with the Rasch or one parameter logistic model (Wright, 1977). The statistics, called person-fit measures can be denoted as U1 and W1, and are derived from standardized residuals. The residual is the difference between observed score and the probability of a correct score at a specific ability level. In the general case the residual is squared and standardized to yield an unweighted fit statistic U1 which can be written as:

$$U1 = \frac{1}{n} \sum_{i=1}^{n} [(U_{ij} - P_{ij})^2] / [P_{ij}(1 - P_{ij})]$$

where U_{ij} is the observed response and P_{ij} is the probability of a correct response and n is the number of items. The second statistic, W1, is a weighted version of U1:

$$W1 = \sum_{i=1}^{n} (U_{ij} - P_{ij})^2 \; / \; \sum_{i=1}^{n} [P_{ij} (1 - P_{ij})]$$

The expected mean of a fit statistic is 1 (Grosse & Wright, 1988; Smith, 1986). These researchers have shown also that the two fit statistics can be standardized to approximately normal (0,1), $U1_S$ and $W1_S$. Smith (1986) and Grosse and Wright (1988) have used $U1_S$ and $W1_S$ in simulation studies; however, those indices have not been generally used in comparative studies.

Three parameter analogs to U1 and W1, U3 and W3 were developed by Rudner (1983) and have also been used by Drasgow, Levine, and McLaughlin (1987) and Harnisch and Tatsuoka (1983). There is no reported use of standardized normal (0,1) versions of U3 and W3.

Extended caution indices. The third general category of IRT based appropriateness indices, extended caution indices (ECI), links the group-dependent and likelihood based (IRT) categories. Several researchers (Tatsuoka, 1984; Tatsuoka & Linn, 1983; Tatsuoka & Tatsuoka, 1982a,1982b) have described the conceptual base for ECI's. In its general form the ECI is an extension of Sato's caution index except that the observed response vector is correlated to a probability vector defined by item response theory. The general ECI can then be written:

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

where $\underline{Y}_i$ = the observed response vector of person i on item j
$\underline{P}_i$ = probability vector associated with person i
$\underline{Y}.$ = the column-sum vector of the observed scores on n items.

In this relationship the numerator is the covariance of an examinee's response vector and the column vector over n items. The denominator is the covariance of the ith row probability from IRT and the column sum vector.

Tatsuoka (1984) identifies six different but related ECI's; however, ECI2 and ECI4 have been found to be the most useful. Expressed in the ratio of two covariances, these indices can be written:

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

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

The indices use the Group Response Curve $\underline{G}$ where each of the elements, G_j, is defined as:

$$G_j = \frac{1}{N} \sum_{i=1}^{N} P_{ij}$$

For a given ability i indexes examinees, j indexes items, n is the number of items, and N is the number of examinees. $\underline{G}$ is the IRT analog for the S curve in Sato's S - P Curve (Tatsuoka, 1984; Tatsuoka & Linn, 1983).

In the case of ECI2 the numerator is the covariance between an examinee observed response pattern and the group response vector as derived from IRT. The denominator is the covariance between the group response vector and the ith examinee's probability vector. ECI4 has the same denominator but the numerator is the covariance between the examinee response pattern and the ith examinee's probability vector. ECI2 purports to identify an examinee whose response pattern deviates from those in the group; while ECI4 is related to the person response curve to which other researchers (Lumsden, 1978; Trabin & Weiss, 1979) have referred. Further, ECI2 is analogous to the NCI and ECI4 is

analagous to the ICI.

Some of the statistical properties of the family of ECI's have been investigated by Tatsuoka and Tatsuoka (1982a) and Tatsuoka (1984). The distributions of the indices were examined by plotting expected values against true scores. The researchers found that the distributions of the indices were U-shaped, with inflated values at high and low scores. To enable comparisons among indices across ability levels the ECI's were standardized by subtracting the expected values and dividing by the standard error in the general form (Tatsuoka & Tatsuoka, 1982a; Tatsuoka, 1984). For ECI2 the conditional expectation is zero and the variance is:

$$\text{var } E\ (ECI2 \mid \theta\ ) = \frac{\sum_{j=1}^{n} \sigma_{ij}^{2}\ (G_j - G)^2}{n^2\ \ \text{cov}^2\ (\underline{G}_i,\ \underline{P}_i)}$$

For ECI4 the conditional expectation is:

$$1\ -\ \frac{\text{Var}\ (\underline{P}_i)}{\text{Cov}\ \ (\underline{G}, P_i)}$$

and the variance is:

$$\text{var } E\ (ECI4 \mid \theta) = \frac{\sum_{j=1}^{n} \sigma^2 ij\ (\ P_{ij} - T_i)^2}{n^2\ \text{cov}^2\ (\ \underline{G},\ \underline{P}_i)}$$

where T_i, the Test Response Curve, is analagous to Sato's P Curve.

ECI2 can then be transformed as ECIZ2 and ECI4 as ECIZ4. The standardized forms, ECIZ2 and ECIZ4, are independent of ability (Tatsuoka, 1984; Tatsuoka & Tatsuoka, 1982a). Further, Tatsuoka (1984) has reported that ECIZ2 and ECIZ4 follow an approximately normal distribution with a mean of 0 and a standard deviation of 1.

The values of the indices are dependent upon the nature of the ratio of the two covariances in the formula. A negative covariance between observed and expected responses indicates that the examinee is not responding appropriately. The ratio would then be a positive number with a larger value indicating a more inappropriate response pattern. If a large positive covariance exists between expected and observed response patterns, a negative value is produced for the index. Tatsuoka (1984) presented an example for a six item test where an inappropriate response pattern produces an ECIZ4 of 1.968 and an appropriate response pattern produces an ECIZ4 of -1.510.

In summary, appropriateness indices based on item response theory offer advantages over purely group dependent measures. They are based on a theoretical framework and are not dependent on item difficulty order. IRT indices are based on relationships between an expected response vector based on ability and an observed vector. Non-standardized forms can be problematic because there may be a curvilinear relationship between the index

and the total score. This is the case if high or low scores produce indices which are more extreme than scores in the average range. The development of standardized forms of the indices may overcome this particular problem. Another proposed advantage of the IRT indices is that the distributions of the standardized indices, including Lz, ECIZ2, and ECIZ4, are considered to be approximately normal (Drasgow, Levine, & McLaughlin, 1987; Tatsuoka, 1984).

Among the disadvantages of IRT based indices is that such indices are more expensive to produce because of the large amounts of computer time required. Application of IRT models also raises questions of model-data fit because the indices are derived from tests which may violate assumptions of unidimensionality, or local independence. In addition, tests may be speeded and response vectors include much missing data. Practical matters such as the need for large samples may also present limitations to the use of IRT based indices.

Research Related to Appropriateness Indices

In this section research related to appropriateness measurement will be presented. Typically, two types of research studies have been used to investigate the characteristics and effectiveness of appropriateness indices. First, studies which used simulation procedures are presented and discussed. Then studies using data from real subjects and studies which can be considered applications of appropriateness indices are presented

and discussed. Finally, a number of problems related to appropriateness measurement are discussed.

Simulation Studies

Typically there are several stages to simulation studies. First, item parameters are either estimated or generated; then, simulated abilities are selected, and non-aberrant response vectors are generated and examinee ability is re-estimated based on the generated response pattern. Aberrant responses are created by changing a given percentage of items in generated response vectors to create either spuriously high or spuriously low scores. An appropriateness index is then computed for both non-aberrant and aberrant samples of response patterns. The effectiveness of the index is determined by assessing the proportion of correct classifications in the aberrant sample for a selected false positive rate. The false positive rate is the probability of incorrectly identifying a truly non-aberrant response pattern as aberrant.

Levine and Rubin (1979) used simulated data to investigate the properties of three indices; Lo, the optimal index, and the square root of the maximum likelihood of the ability variance. Item parameters, estimated from the 85 item Scholastic Aptitude Test-Verbal (SAT-V) were used to generate response vectors for 2800 simulated non-aberrant examinees. These were simulated responses of examinees assumed to be responding appropriately. Eight combinations of aberrant response patterns based on two types of aberrance, spuriously high and spuriously low and four

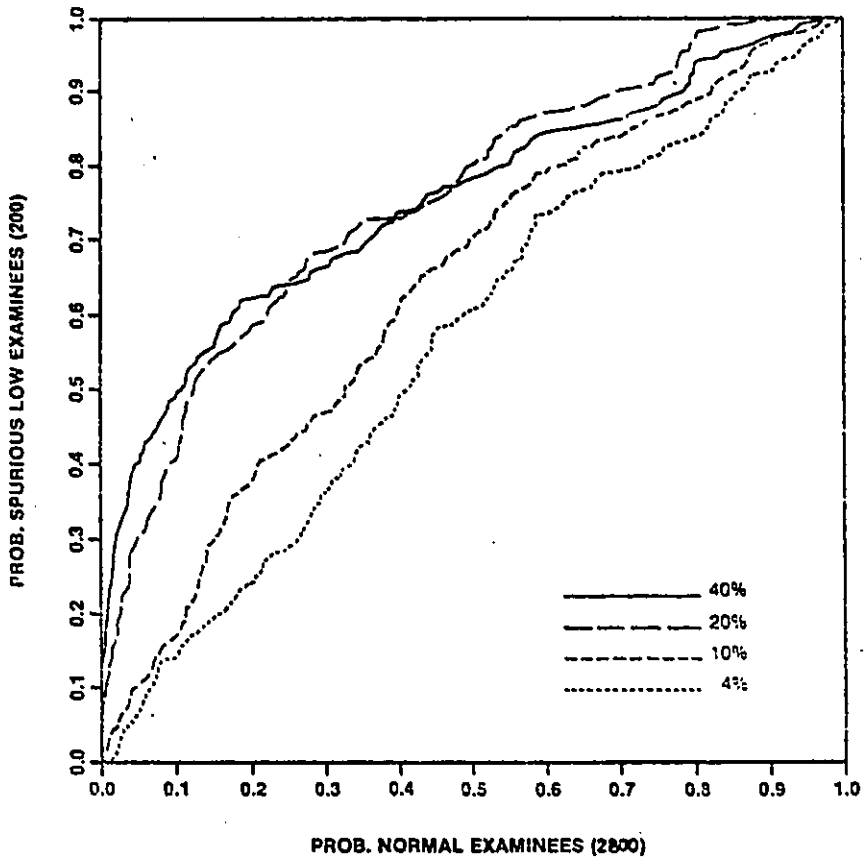
levels of aberrance, 4%, 10%, 20%, 40%, were established. For each of the eight combinations, 200 simulated responses were generated.

Spuriously high scores were created by changing a percentage of the incorrect to correct scores. For example, 4% aberrance was created by changing 4% of the item scores from incorrect to correct. As a result of such tampering, the response patterns would over-estimate the score with respect to examinee ability. For a given percentage of items spuriously low scores were created by rescoring correct items based on the probability of a correct score under guessing conditions. For example, a correct five option item would be re-scored .20 to simulate guessing. In this way the response pattern produces a score which under-estimates the examinee score with respect to the examinee's ability.

To determine the effectiveness of the indices, Levine and Rubin (1979) used the receiver operating curve (ROC). When using the ROC, appropriateness indices are first calculated for samples of both non-modified responses and the aberrant responses. Many cutting scores are then established and the proportions of each sample exceeding the cutting score at each value are plotted as points on a graph. A large number of cutting scores would produce curves such as in Figure 3 which shows an example of the ROC for each of four levels of aberrance with a sample of spuriously low examinees. The x axis can be considered as the false positive rate. Depending upon what is judged to be an

Figure 3.

ROC curves for Lo with four levels of spuriously low scores.



Note. M.V.Levine and D.B.Rubin " Measuring the Appropriateness of Multiple-Choice Test Scores" Journal of Educational Statistics, 4, p.279. Copyright 1979, American Educational Research Association, Washington, D.C. Reprinted by Permission.

acceptable false positive rate, the ROC can be used to identify the proportion of correctly classified aberrant response patterns. If the y axis represents the detection rate and the x axis is the false positive rate, it can be seen that a desirable index would produce an ROC which is close to the y axis (i.e a high detection rate for a low false positive rate). Figure 3

also shows that the detection rate is related to the degree of aberrance in that under higher aberrancy rates more aberrant response patterns are detected.

Results of the study showed that the detection rates of each of the three indices were similar and that spuriously high scores tended to be more easily detectable than were spuriously low scores. Detection rates also tended to improve as the level of aberrance increased and spuriously high scores tended to produce higher detection rates than spuriously low scores. For example, at 40% spuriously high scores and .10 false positive rate about 70% to 80% detection rates were achieved. At 40% spuriously low and a .10 false positive rate, detection rates were about 45% to 60%. Similar patterns of detection rates though with lower detection rates were reported for both 4% and 20% aberrance. Spuriously high scores also produced higher detection rates than spuriously low scores for these lower levels of aberrance. This may be explained as a consequence of the method of generating aberrant response patterns. The spuriously low scores were created by replacing the correct score with a probability of guessing. On the other hand spuriously high scores were created by re-scoring incorrect items to correct which results in more changes in a response vector than is the case of spuriously low scores.

Levine and Drasgow (1982) used simulated data from Levine and Rubin (1979) to investigate the original, unstandardized Lo-type indices with spuriously low response patterns on the SAT-V.

To investigate whether aberrant responses in the norming sample degraded the indices effectiveness, 2800 approximately normal responses and the 200 responses at 20% spuriously low were merged. Item parameters were estimated using the heterogeneous sample, and Lo was used to measure the aberrant response patterns. The results were compared with those obtained when the item parameters were estimated exactly from the norming sample. The detection rates for the two samples were not noticeably different; this was taken to mean that a relatively high proportion of aberrant responses such as found here would not degrade the effectiveness of the index. Levine and Drasgow (1982) also found that the index could be used with either estimated or simulated item parameters with little differences in results.

Levine and Drasgow (1982) investigated Lo using item parameters from the 95 item Graduate Record Exam-Verbal (GRE-V). Responses from 10,000 examinees were used to estimate the item parameters, and 2270 response vectors were selected as the approximately normal sample composed of examinees considered to be responding appropriately. Another 200 response patterns were subjected to the 20% spuriously low treatment and detection rates were computed for Lo. Results showed a detection rate of about 40% at a false positive rate of .037 which was similar to that found for the SAT-V.

A problem in appropriateness measurement is the extent to which the IRT model used for estimating the item parameters has

an effect on the index. Drasgow (1982) examined the effects of using both the 3PLM and the Rasch model to estimate item parameters. The indices used were Perbis and the three indices suggested by Levine and Rubin (1979) except that Lg was used rather than Lo. Responses from 10,000 examinees on the GRE-V were used to estimate item parameters and 5,000 respondents were selected as the nominally normal sample. A test norming sample (N = 3000) was used to establish the non-aberrant conditions. To examine the relationship of ability to the indices two ability groups, a low ability group (N=115) and a high ability group (N=200), were identified. Spuriously low scores were produced for the high ability group and spuriously high scores for the low ability group. In each case 20 of the 95 items were randomly selected for modification. Detection rates for the four indices were compared for estimates derived from each of the 3PLM and the Rasch model.

Results showed little difference in detection rates among the three IRT indices at most false positive rates for either spuriously high or spuriously low scores. For example, at a .05 false positive rate, the three IRT indices identified 46% to 50% of the spuriously low responses under the Rasch model and 50% to 58% using the 3PLM.

Another test norming sample (N = 500) was used to examine the effects in smaller samples. Results similar to those for the larger sample were produced. This suggests that if computation time were a concern, a one parameter model might be preferred in

appropriateness measurement. A limitation to this study is that the non-standardized indices used may have contributed to some of the variability because unstandardized indices have been found to be dependent on ability.

Drasgow (1982) also reported that spuriously low scores were more detectable than were spuriously high scores for the GRE-V. For example, at a .05 false positive rate about 50% of spuriously low scores were detectable but less than 20% of the spuriously high scores were detectable. These results differ from those of Levine and Rubin (1979) who for the SAT-V found spuriously high scores were more detectable. A possible explanation is that in contrast to the GRE-V, the SAT-V has many items with a small value for the "c" parameter; thus low ability examinees are unlikely to get difficult items correct by guessing. If the procedure for modifying response patterns then assigns correct response to these examinees, they become more conspicuous and more detectable by Lo.

Rudner (1983) used simulation methods to evaluate a total of nine group-dependent and IRT based indices including NCI, MCI , PERBIS, the personal- point biserial, U1, W1, U3, W3, and Lg. Data were generated to simulate two types of tests, 80 items of the norm-referenced SAT-V and a 45 item classroom general biology test. Item parameters were specified from previous parameterization of the tests. To establish the non-aberrant condition, 2000 simulated examinees were randomly generated from a normal distribution of abilities. These simulated responses

are assumed to be those of examinees who are responding appropriately. Simulated response patterns were generated for 2000 examinees in each of the two tests.

Unusual response patterns were created for each of eight samples of 100 by altering randomly selected items to generate both spuriously high and spuriously low scores for examinees for 5%, 10%, 15%, and 20% for the SAT-V and 5%, 10%, and 15% aberrance for the general biology test. Consequently, there was a sample of 2000 un-modified response vectors and a total of eight sets of 100 modified response vectors, four spuriously high sets and four spuriously low sets, for the longer test. For the shorter test there were six sets of 100 modified response patterns plus the sample of 2000. Indices were calculated for the spuriously high and spuriously low scores for each of the two tests.

To assess the effectiveness of the indices detection rates based on a .05 false positive rate were computed for each index for each type of test for each type of aberrance. No consistent pattern of detection rates were found. Some indices produced higher detection rates for some conditions and other indices for other conditions. Further, there appeared to be interactions between type of test, type of aberrance, and level of aberrance which affected detection rates. The highest detection rates, 50% to 75% were produced by W3, U3, and Lg for the SAT-V test for 15% and 20% spuriously high scores. Generally the IRT indices produced higher detection rates; however, for spuriously high

scores for the biology test MCI, NCI, and PERBIS produced slightly higher detection rates.

In general spuriously high conditions resulted in slightly higher detection rates although at 5% aberrance spuriously low scores produced higher detection rates for U3, W1, and U1. Also the SAT-V tended to produce higher detection rates than the biology test although at 15% aberrance the biology test produced higher detection rates for U1, W1, and Lg. As well higher levels of aberrance generally produced higher detection rates than lower levels of aberrance. However, for 10% spuriously low scores U1, U3, and W1 produced lower detection rates than did 5% spuriously low for both the SAT-V and biology tests.

In comparing the IRT based indices to the group-dependent indices, the researcher reported that the IRT based indices generally produced higher detection rates for most conditions. However, MCI and NCI were more consistent in that they identified similar proportions of aberrant response patterns in all situations.

Rudner also reported the correlations among the indices for each of the two types of tests. The four group-dependent indices correlations were distributed in a range from -.99 to .95 for the SAT-V and from -.99 to .97 for the biology test. Correlations among the IRT indices for the SAT-V ranged from -.80 to .74. Those indices based on the 3PLM (U3, W3, and Lg) produced the highest correlations. Correlations between the categories of indices, group-dependent versus IRT based, tended to be lower

than correlations within the categories of indices. In contrast to the SAT-V, correlations among the 3PLM indices were relatively low, -.39 to .33, for the biology test. In general, however, correlations for the SAT-V tended to be higher than those for the biology test.

Parsons (1983) explored the use of Lg as an appropriateness index with the Job Descriptive Index (JDI), an instrument that requires typical responses rather than measuring maximum response. The JDI normally requires a three option answer, yes, no, or don't know. The researcher modified the scale to create a 1,0 pattern by collapsing the latter two categories. Using a sample of 1906 workers, LOGIST was used to estimate difficulty and discrimination. Using these parameters and three levels of ability,-2 ,-1, and 2, three datasets were generated. Because the original questionnaire had used a three option response, yes, no, don';t know, it was determined that the latter two categories could be considered as one since don't know was judged to be akin to no. A sample (N = 1000) of the original un-modified records were selected as the non-aberrant response patterns. These were then used to establish detection rates using the ROC.

Aberrant response patterns were created by modifying one third of the items of each response vector in each sample. Because of the nature of the original questionnaire, 0's were changed to 1's with a probability of .33 and 1's to 0's with a probability of .66. After computing Lg for each simulated respondent, detection rates were produced using a ROC in each

sample for both the total 60 item test and for an 18 item subtest of the JDI. Results showed that the detection rates were better for the longer test and for the spuriously low scores of the high ability group. At a false positive rate of .10 about 60% of the aberrant response patterns were detected for the high ability group. One of the problems in research such as that conducted by Parsons is that the extent of aberrancy in the non-aberrant response patterns is not really known and therefore the accuracy of the detection rates might be in question.

Drasgow, Levine, and Williams (1985) conducted studies to examine three aspects of appropriateness measurement: the effect of omit rates, the use of polychotomous response models, and standardization of the Lo indices. Lz was developed as a standardized index for dichotomous scoring and Zh for polychotomous scoring. A sample of 3000 was drawn from a large number of respondents to the 1975 SAT-V and LOGIST was used to estimate item parameters and ability levels. These parameters were used to examine the pattern of option choices for about fifty thousand SAT-V respondents in order to establish 25 categories of abilities ranging from -2.05 to 2.05. The researchers first investigated the distribution of the indices in a nominally normal sample. These were the response patterns considered to be non-aberrant. Lo, Lz, and Zh were calculated from a sample (N = 464) of examinees from a wide range of abilities. When the three indices Lo, Lz, and Zh were plotted against ability, Lo was found to be dependent upon ability to a

much greater extent than were Lz and Zh. The conditional distribution of Lo varied linearly as a function of examinee ability, as a result it was difficult to interpret the index directly. Of the three indices, Lz was least dependent upon examinee ability.

Using Lz and Zh with a larger sample (N= 3478) it was found that the indices were distributed similarly across different ability levels, although they were not standard normal. The shape of the distributions of each index was compared by examining the cumulative proportions of index scores at various cut-off scores across ability levels. The proportion at each ability level was compared with the proportion expected for a normal curve at each cutting score. If there were similar proportions of the index scores at various ability levels, it was assumed that the index was not dependent upon ability. Although the indices were not statistically independent of ability, the effects were small and did not appear to degrade the value of the indices.

For this investigation, four false positive rates (.005, .025, .050, and .097) were used with the following associated cut-off scores for both Lz and Zh: -2.58, -1.96, -1.64, and -1.30. Detection rates for Lz and Zh were investigated by using the statistics produced from the 3478 nominally normal, non-aberrant, examinees. Samples (N= 300) of spuriously high and spuriously low response patterns with three levels of aberrance 10%, 20%, and 30% were produced by modifying previously generated

response patterns. Procedures to modify response vectors and to compute detection rates of indices were similar to those used by Levine and Rubin (1979).

Detection rates for both spuriously low and spuriously high aberrance were found to be related to examinee ability and to the degree of aberrance. For example, at a false positive rate of .10, spuriously low scores for high ability examinees produced a detection rate of about 90% for both Lz and Zh at 30% aberrance. At 10% aberrance detection rates declined to about 65%. Spuriously low scores for low ability examinees were only about 40% detectable by Lz at each level of aberrance and 50% to 60% detectable by Zh.

Spuriously high scores were more detectable for low ability examinees with Lz achieving a rate of about 90% at .10 false positive rate and 30% aberrance. Under these same conditions, spuriously high scores were only about 45% detectable in high ability examinees. In contrast to spuriously low scores, Zh performed somewhat less well than Lz for spuriously high scores at all levels of ability and for all levels of aberrance. These results are expected insofar as spuriously high scores are more detectable in low ability examinees and spuriously low scores in high ability examinees. For example, it is difficult to modify an already low score to appear spuriously low since the resultant response pattern may be quite similar to that of the original.

Drasgow, Levine, and McLaughlin (1987) investigated several other aspects of appropriateness measures in a series of studies.

In the first of these studies eleven indices were compared on the basis of standardization and power in a series of simulations. The indices included the following: LR, Lz, U3, W3, ECIZ2, ECIZ4, MCI, a jackknife technique, item-option variance, and two observed/expected information ratios. These latter four indices were suggested by the researchers based on the notion that inappropriate responses will flatten the likelihood function near its maximum rather than have a sharp maximum. With respect to standardization, an index was considered well standardized if the conditional distribution of the index was approximately equal across ability levels for non-aberrant response patterns.

To investigate standardization, item parameters and abilities were estimated for the 85 items on the 1975 SAT-V from about 50,000 examinees. Five categories of ability, low through high, were established and 3000 examinees were simulated in each category. These categories were considered to include examinees each of whom are responding appropriately. Each of the indices was computed for each of the five samples and compared using ROC, with each of the other four samples. In this study the Lz, W3, jackknife, ECIZ2, and ECIZ4 were shown to be fairly well standardized.

To investigate the detection rates of the indices, data were generated from the same SAT-V data for 4000 simulated unmodified response patterns and for 2000 simulated aberrant response patterns for each of twelve conditions. Aberrance was created at two levels 15% and 30% for each of simulated spuriously high or

low response vectors, and for six categories of ability. The full range of abilities -2.05 to 2.05 was divided into six categories; three were levels of high ability, for example, very high, high, and high average, and three were levels of low ability. The three high ability groups were used to create spuriously low scores and the three low ability groups used to create spuriously high scores. This was done because it was assumed that it was more important to detect spuriously low scores in high ability examinees than in low ability examinees and to detect spuriously high scores in low ability examinees.

To examine detection rates of the indices each of the twelve simulated aberrant samples was compared with the normal or unmodified sample. The results showed a wide range of detection rates. The detection rates ranged from zero to over 90% depending upon the index and degree of aberrance. Of the eleven practical indices which were evaluated for standardization and effectiveness, Lz, ECIZ2, ECIZ4, and W3 tended to produce the best detection rates but effectiveness varied somewhat with type level of aberrance. For example, for very low ability and 15% spuriously high Lz identified 69% of the aberrant scores at .05 false positive rates; for 30% spuriously high the detection rate was 96%. The detection rate for spuriously high scores was higher for very low ability examinees than for low average examinees and similarly spuriously low scores were more detectable for very high ability examinees than for high average examinees. This is not surprising because spuriously low scores

appear more aberrant for higher ability examinees. At a .05 false positive rate over the two types and two levels of aberrance, Lz correctly identified 35% to 98% of the modified scores; W3 identified 29% to 98%; ECIZ2 identified 32% to 98%; and ECIZ4 identified 29% to 97%. For the non-standardized indices, detection rates ranged from 1% to 80% for item-option variance, 18% to 91% for the observed/expected likelihood curvature and 23% to 99% for the jackknife procedure. The wide range of detection rates may in part be due to the fact that examinees were grouped by ability. Spuriously low scores tended to be more detectable in high ability examinees and spuriously high scores tend to be more detectable in low ability examinees.

In a follow up study to examine the effects of shorter tests on the indices, Drasgow, Levine, and McLaughlin (1987) evaluated eleven indices using data from the Armed Services Vocational Aptitude Battery (ASVAB). Two studies using simulated data and one using actual data were conducted. Two subtests, the 30 item Arithmetic Reasoning (AR) test and a combined 50 item Word Knowledge/Paragraph Comprehension (WK/PC) test were used.

In the two simulation studies LOGIST was used to estimate the item parameters from a sample of 2978 examinees. Using these parameters a sample of 4000 non-aberrant response vectors was generated. Aberrancy was produced in sixteen samples of 2000 generated response vectors. Spuriously high scores were created in each of four ranges of low ability examinees (very low to low average) at each of two levels of aberrance, 15% and 30%.

Similarly, spuriously low scores were created in a total of eight samples of high ability examinees which included four ranges of ability (high average to very high) and two levels of aberrance. Appropriateness indices were computed for each examinee in each of the sixteen samples.

The results for two tests, a 30 item AR and a 50 item WK/PC, showed that detection rates were generally higher for the WK/PC than for the AR. Overall, Lz, Zh, ECIZ2, and ECIZ4 were the most effective even though those indices failed to detect 50% of the simulated aberrant scores in low average or high average ability ranges at a .05 false positive rate. Spuriously low scores for high ability examinees were more detectable than were spuriously high scores for low ability examinees. For WK/PC at a .05 false positive rate and 30% spuriously low, about 90% of the modified scores were detected by Lz in the very high ability range but only 61% in the high average ability range. With the shorter AR, detection rates were 76% in very high abilities and 30% in high average abilities. For spuriously high scores with the WK/PC, about 76% were identified in very low ability examinees with 30% aberrance and about 20% in average ability examinees. Detection rates for AR were 62% and 35% for very low ability and low average ability examinees respectively.

Although these results suggest that detection rates are lower with shorter tests, it should be noted that two different tests were used. Different tests have different item parameters which are used in the computation of the indices. Therefore, the

indices may have been affected by either test type, test length, or a combination of test type and test length.

To examine the effectiveness of the indices in actual ASVAB data, a follow-up study was conducted using the AR and WK/PC results from the National Research Opinion Centre. A norming sample of 2716 actual response vectors was selected and used to estimate item parameters. Then 5428 other examinees' response vectors were sorted into six ability groups from very high to very low. Spuriously high and spuriously low response vectors were created by changing responses of 15% and 30% of the items using the same techniques as described for previous studies. Detection rates were calculated for LR, Zh, Lz, ECIZ2, ECIZ4, U3,and W3 indices by using the false positive rates from .001 to .10.

With respect to detection rates for WK/PC, the results for the real data were similar to the simulated ASVAB data; Lz, ECIZ2, and ECIZ4 were most effective in detecting unusual responses. Lz was found to produce the highest detection scores at lower ability ranges. For moderately low ability examinees detection rates for Lz were about 45% at a false positive rate of .05 with 30% spuriously high aberrance. Rates for ECIZ2 and ECIZ4 were also about 45%, for W3 about 25%, and for U3, 10%.

For spuriously low scores, detection rates were not as high as with the simulated data. With real data at moderately high ability levels, 30% aberrance, and .05 false positive rate, the detection rates were about 80% for Lz, ECIZ2, and ECIZ4 and about

65% and 55% for W3 and U3 respectively. The other indices used in the study failed to produce detection rates as high as the three most effective ones. The researchers reported that the distribution of examinees in the norming sample had more extreme values for real data than for the simulated data. The real data showed a large proportion of subjects with very low abilities, which was offered by the researcher as an explanation as to why the spuriously high scores were more detectable.

With respect to fit statistics as appropriateness indices, Rudner (1983) and Drasgow, Levine, and McLaughlin (1987) as described earlier, have investigated the 3PLM statistics, U3 and W3. Smith (1986) has investigated the use of a Rasch model fit statistics, similar to U1, to detect unusual response patterns. Data were simulated for a total of 1900 examinees for a ten item test. At a false positive rate of .05 it was found that the index produced a detection rate of 50% for 33% spuriously high scores. Random guessing produced a detection rate of about 70%. Although the distribution of the index was not known, the researcher suggested that values greater than 2.00 would occur less than 5% of the time.

In summary, simulation studies have been used with appropriateness indices to assess their effectiveness in detecting aberrant response patterns. Such studies have been most useful in the investigation of IRT based indices. Simulation studies are of particular interest because they enable the researcher to establish a norming sample which is considered

to be responding appropriately. In addition, the extent of aberrance is actually known. Simulation studies also allow the researcher to replicate a study many times over, although this procedure has not yet been used specifically in appropriateness measurement.

Some general conclusions can also be drawn with respect to the detection rates of appropriateness indices. Results of more recent research, (Drasgow, Levine, & Mclaughlin, 1987; Drasgow, Levine, & Williams, 1985) suggest that standardized indices produce consistently higher detection rates than do non-standardized indices. These results also indicate that higher levels of aberrance result in higher detection rates than do lower levels of aberrance.

With respect to type of aberrance, Drasgow, Levine, and McLaughlin (1987) have indicated that spuriously low scores tend to produce somewhat higher detection rates than spuriously high scores. However, recent researchers have used categories of ability to assess detection rates. As expected, effectiveness of indices in detecting spuriously high scores is higher for low ability examinees. Similarly, spuriously low scores are more easily detected for high ability examinees. Rudner (1983) reported that spuriously high scores were somewhat more detectable than spuriously low scores. However, it must be noted that different modification procedures were used by Rudner than were used by Drasgow, Levine, and McLaughlin.

Results of research with respect to effect of test length

are also somewhat inconsistent. Rudner (1983) and Drasgow, Levine, and McLaughlin (1987) suggested that longer tests produced higher detection rates than shorter tests; however, in those research studies the extent to which test type was confounded with test length is not known. It is not known whether tests of different lengths, but derived from similar item parameters, would differentially affect the indices.

Results of research comparing detection rates of different indices are somewhat inconclusive and contradictory. For example, Drasgow, Levine, and McLaughlin (1987) reported that Lz produces higher rates as compared to other indices at low levels of aberrance and for spuriously high scores. W3 and U3 are more effective at lower levels of aberrance and for spuriously low scores. ECIZ2 and ECIZ4 appear to perform better at higher levels of aberrance for both spuriously high and spuriously low scores. Overall in simulation studies IRT indices were more consistent and effective than group-dependent indices, although the MCI was nearly as effective in some cases (Rudner, 1983). However, it is also possible that the simulation procedures which use the 3PLM may bias detection rates in favour of IRT indices as compared to group-dependent indices which are not based on IRT. In only one study was the effect of IRT model examined and in that case the Rasch and 3PLM were compared with unstandardized indices only.

Although simulation procedures provide a number of advantages, one disadvantage is that such procedures may also be

costly as well as requiring large samples. Further, by definition real subjects are not used; therefore it is not known whether the results are readily generalizable to real testing situations.

Studies Using Data from Real Subjects

In this section studies which have employed appropriateness indices with data from real subjects are reviewed. Of particular interest are results of studies in which researchers have used both IRT based and group-dependent indices. A number of researchers including Harnisch and Linn (1982), Harnisch (1983), Koffler (1983), Oltman (1985), Jaeger and Busch (1986), and Miller (1986) have used group-dependent indices only and will not be reviewed here.

Frary (1982) used the U1, W1, and the MCI with classroom test data from about 2000 college students from twelve classes. The tests were developed and administered by different professors and were typical of college level multiple choice tests. The three indices were calculated for each student who took the tests. Correlations with total score ranged from 0 to -.36 with MCI and W1 showing lower correlations than U1. Intercorrelations among the three indices ranged from .42 to .88 across the twelve classes. The correlations with total test score were reported to be less than $\pm$.50 over the twelve tests. As reported in several other studies using real data, no assumptions were made about the prior probability of aberrant response patterns and no detection rates were computed. Consequently, the extent to which aberrant

patterns exist is not known in data for real subjects.

Harnisch and Tatsuoka (1983) conducted a similar investigation to evaluate various properties of fourteen IRT based appropriateness indices. The indices were ECI1, ECI2, ECI3, ECI4, ECIZ1, ECIZ2, ECIZ3, ECIZ4, U1, W1, U3, Lo, Ln, and Lz. Data were obtained from a representative sample of 2437 thirteen year old students who had written the National Assessment of Educational Progress mathematics test. Results showed that, except for U1, the indices were highly intercorrelated (>.70). As may be expected the highest intercorrelations were among the standardized and unstandardized ECI's (>.80).

Correlations with the total test score produced inconsistent results. Although most of the correlations were less than .10, four indices, Ln, Lz, W1, and U3, produced correlations with the total score in the range -.23 to .36 with Ln being the largest and the other indices about + or - .25. Using the change in R squared as the test of curvilinearity, the researchers reported that of the 14 indices, only W1 and Lo showed a curvilinear relationship with total score; therefore ability should be controlled when using those two indices. Further, the researchers observed that only standardized indices should be used as appropriateness measures because such indices seemed to show the least dependence upon total test score. Thus, the group standardized extended caution indices were seen as the most promising appropriateness measures.

Birenbaum (1985) examined the effectiveness of nine IRT-based measures in distinguishing unusual response patterns of 310 students on an English as a second language comprehension test. The IRT based indices investigated in the study were ECI1, ECI2, ECI4, ECIZ1, ECIZ2, ECIZ4, Lo, Lz, and U1. The sample included both real respondents (N=280) and simulated responses (N=30). For the purpose of establishing detection rates 238 response patterns were defined as appropriate and the response patterns of 72 respondents including the 30 simulated responses, were a priori defined as inappropriate. This classification was based on a description of the students as cooperative or uncooperative. The purpose of the study was to detect the uncooperative examinees who produced the inappropriate response patterns. The cut-off point for classifying an index score as inappropriate was the upper 97.5% point of the distribution of the scores of the 238 cooperative examinees who were considered to be appropriately responding. Item parameters were estimated using the 2PLM.

Results of the study showed that intercorrelations among the indices ranged, in absolute value, from .25 to .98 among ECI1, ECI2, ECI4, ECIZ1, ECIZ2, ECIZ4, Lz, and U1. Correlations among the ECI's and ECIZ's ranged from .78 to.91. For Lo, the correlations ranged from -.31 with ECI1 to -.72 with ECIZ2. Lo produced a correlation of .75 with Lz and -.34 with U1. Correlation between U1 and the ECI's and the ECIZ's were relatively low with a range of .310 to .441. U1 and Lz produced a correlation of -.454. Finally, Lz produced the most

consistently high correlation with the other indices. Correlations ranged from -.788 to -.486 with the ECI's and ECIZ's. Lo, the unstandardized index produced lowest correlations (.007) with the total score and ECIZ4 the greatest correlation with total score (-.233). ECIZ2, ECIZ4, and Lz were found to be the most effective with detection rates of 73% to 83%; however, to achieve those detection rates, a false positive rate of about 20% was used.

Chatman (1985) investigated the use of the ECIZ4 to predict college mathematics achievement. The researcher assumed that information on unusual response patterns would be helpful in making decisions regarding placement of first year college math students in various types of courses. It was assumed that ability of a student with an inconsistent (unusual) response pattern on an entrance test may be inaccurately estimated by the test. Data were collected for each of two successive years on about 700 students to examine pre-course math scores as a predictor for subsequent achievement. The results of the study indicated that there was some relationship between the ECIZ4 value and mathematics ability, but generally the index did not contribute enough additional information to be useful in placing first year math students.

Nelson and Chatman (1986) used the ECIZ4 and U1 in a study of self-reported guessing with 207 college students on a psychology test. The two indices were quite highly correlated, .74, but there was virtually no relationship between self-

reported guessing and each of the two indices. It had been assumed that guessing was more common among low scoring examinees; however, results showed no difference between guessing by high scoring and guessing by low scoring subjects. These results may be interpreted to mean that the indices were not useful in identifying guessing as is purported, although this interpretation was not offered by the authors.

Tomsic (1986) examined the distribution and stability of standardized and unstandardized extended caution indices with school achievement test data. Data were used from two samples, (a total of about 1400) elementary school students who were administered the Comprehensive Tests of Basic Skills on two occasions one year apart. The results indicated that the indices showed little relationship with total test score. Distributions of the indices were stable over various sample sizes and between gender. It was reported that the distributions of the indices deviated from normality as measured by goodness of fit tests. This finding was of interest since it would therefore be difficult to compare the index across samples. However, since this study involved data for real subjects, it was not possible to know the extent to which aberrance may or may not have been present in the data. Further, the study was based on the assumption that an examinee who produced an aberrant response pattern in the first year of the testing, would also do so in the second year. Correlations for each index over the two administration periods were somewhat low, from .05 to .30,

suggesting that the indices were not a consistent measure for individuals over test periods. This may be explained because many factors may cause a student's score to be aberrant on one test period and not so in a subsequent test period. In effect, one cannot assume aberrancy is consistent from year to year for a student.

In summary, results of studies using real data have provided evidence with respect to the intercorrelations among indices. One advantage of such studies is that the results may be generalized to similar situations. Real data studies also allow the researcher to make practical comparisons among indices. Because small samples are often used, computation of indices may be relatively cost effective. An important limitation of studies using real data is that since the extent of aberrance in a sample is not known a priori, the true effectiveness of an index can not be ascertained.

One of the benefits of studies using real data is that correlations among indices may be examined. One of the most consistent results is that standardized IRT-based indices tended to produce somewhat higher intercorrelations and less evidence of dependence upon total test score. Overall, it appeared that the family of ECIZ's produced the highest intercorrelations (>.80). Lz also seemed to be highly correlated with the ECIZ's and U3. U1 and W1 produced lower intercorrelations with ECIZ's, with absolute value of about .40.

Summary of Research

Researchers who have investigated appropriateness measurement have used both simulated and real data in addressing the effectiveness of, and the relationship among, various appropriateness indices. They have studied the effects of types of aberrance, levels of aberrance, test length, and IRT model. In addition, they have examined the correlations among the indices. In this section these studies are reviewed and some conclusions drawn from the results of the research. Also a number of problems which need to be addressed in future research are identified.

Simulation studies and studies using data from real subjects are the two general types of research methods used in appropriateness measurement. Simulation studies have been particularly useful in providing evidence with respect to the detection of aberrant response patterns by indices. This is possible because in the simulation studies a norming group, which can be considered to be responding appropriately, can be used to establish false positive rates. Thus it is possible to examine the power of a statistic to correctly classify simulated aberrant response patterns at a specified false positive rate.

Simulation procedures also enable the researcher to replicate an experiment many times which may be desirable in studies of the stability of distributions of statistics over samples. Although data from actual subjects have the advantage of being real data, it is not possible to compute accurate

detection rates because the extent of aberrance in the data is not known. Simulation studies may be criticized since by definition they do not use data from live subjects and consequently results may be difficult to generalize. Overall it can be concluded that a simulation study is preferable to a study using real data in situations where the effectiveness of an appropriateness index is being investigated. One never knows the true detection rates with real data.

An important concern in evaluating the effectiveness of an index is standardization. Standardization of an index reduces dependence of the index on examinee ability and total test scores. Of the two general types of indices, only the IRT based can be standardized (Drasgow, Levine, & Williams, 1985; Tatsuoka & Tatsuoka, 1982a). Standardization of an IRT index is based on the conditional expectation of the index, given ability, and therefore it is necessary to estimate examinee ability in order to standardize an index. This is possible with IRT based indices because such indices are derived using estimated examinee ability.

Drasgow, Levine, and McLaughlin (1987) conducted the only reported study which compared a number of indices and found that standardized appropriateness indices are less dependent upon ability and total score and produced higher detection rates than non-standardized indices. Harnisch and Tatsuoka (1983) suggested the standardized extended caution indices were the most promising because they had lower correlations with total test

score. Frary (1982) did not use a maximum likelihood estimate index or extended caution indices but did report low correlations between the appropriateness indices and total test score. Of the IRT based indices Lz, ECIZ2 and ECIZ4 are reported to have an approximately normal (0,1) distribution. Tatsuoka and Tatsuoka (1982a) have demonstrated that although ECI2 and ECI4 had a curvilinear relationship with the total score, this was not true of ECIZ2 and ECIZ4. Similarly, Drasgow, Levine, and Williams (1985) showed that Lo was dependent upon total score to a greater extent than was Lz, the standardized form of the index.

Drasgow, Levine and Mclaughlin (1987) have investigated standardization with respect to the distribution of an index across different levels of ability. In principle, if the distribution of an index is similar across ability levels, the index is said to be well standardized under conditions of no aberrance. Those researchers found that U3 and W3 to be well standardized which suggests that those two indices may perform in a manner similar to other standardized indices such as Lz, ECIZ2 and ECIZ4.

The effectiveness of an appropriateness index is determined by the detection rate which is measure of the correct classification of an aberrant response pattern. Drasgow, Levine, and McLaughlin (1987) reported that Lz, U3, W3, ECIZ2, and ECIZ4 produced the highest detection rates over type of aberrance, level of aberrance at false positive rates of from .001 to .10. At .10 false positive rate, approximately 90% of the aberrant

response patterns were correctly classified by those indices for both spuriously high and spuriously low scores at 15% aberrance. Rudner (1983) who did not study any of the ECI's, ECIZ's, or Lz, found U1, W1, U3, and W3 to be more effective than Lg or any group-dependent indices. Birenbaum (1985) attempted to examine detection rates using data from real subjects. Detection rates for Lo, Lz, the ECI's, and ECIZ's were similar (from 68% to 83%), however, the false positive rates were high, ranging from 16% to 31%. U1 was the least effective index. In general the standardized indices, Lz, ECIZ2, and ECIZ4 provided the highest detection rates; however, U3 and W3 were also shown to be effective in some situations.

Information on the characteristics of the distributions of the indices in non-aberrant conditions is important because detection rates are based on cut-off scores derived from non-aberrant response patterns at given false positive rates. Tatsuoka and Tatsuoka (1982b) have suggested that the distribution of ECIZ2 and ECIZ4 are approximately normal; however, that conclusion was based on a small sample using real data and therefore ma· not be generalizable to other situations. Drasgow, Levine, and McLaughlin (1987) reported the mean and standard deviation of ECIZ4, Lz, and W3 using simulated data (N = 1000). The mean and standard deviation of ECIZ4 was -.14 and .86 respectively. For Lz the values were .09 and .97, while for W3 the values were .99 and .12 for the mean and standard deviation. As expected the means for Lz and ECIZ4 were near 0 and near 1 for

the index W3.

Drasgow, Levine, and Williams (1985) examined the left tail of the distribution of Lz using simulated data (N = 3478). The results of the analysis showed variability in the tail of the distribution, although the researchers concluded the distribution was near normal. No similar investigation of the characteristics of the distributions has been reported for other appropriateness indices. It can be concluded therefore, that further research is needed on the characteristics of the distributions.

Detection rates have typically been computed based on the proportion of items which are considered aberrant in an examinee's response pattern. It has been found that detection rates varied according to type and level of aberrance. Levine and Rubin (1979) found spuriously high scores more detectable than spuriously low scores and Rudner (1983) found spuriously high slightly more detectable than spuriously low. Drasgow (1982) and Levine and Drasgow (1982) reported spuriously low scores more detectable; however those researchers used non-standardized indices only. More recently, Drasgow, Levine, and Williams (1985) found Lz more effective than Zh with spuriously high scores and the opposite for spuriously low scores. The difference in results for these two standardized indices was attributed to the fact that Zh is based on the pattern of option choices of an examinee, while Lz was based on dichotomous scoring. In a recent study in which several standardized indices were compared, Drasgow, Levine, and McLaughlin (1987) reported

detection rates similar for both spuriously high and spuriously low scores. It can be concluded then, that further research is needed to examine detection rates of spuriously low and spuriously high scores.

The effect of the level of aberrance in modified response patterns has been reported by Levine and Rubin (1979), Rudner (1983), Drasgow, Levine, and Williams (1985), and Drasgow, Levine, and McLaughlin (1987). Detection rates tended to improve as the level of aberrance increased, although one study by Rudner (1983) produced higher detection rates at lower levels of aberrance than at higher aberrance. It is not clear, therefore, whether a systematic relationship exists between levels of aberrance and detection rates of indices. It can be concluded therefore that further research is needed on the relationship between level of aberrance and the detection rates of indices.

The effect of test length on the distributions and effectiveness of indices has not been well researched. As discussed earlier, Rudner (1983) found differences between detection rates when using an 80 item SAT-V test and a 45 item classroom biology test. Using non-standardized indices, detection rates were about 10% to 15% higher for the SAT-V test. Drasgow, Levine, and McLaughlin (1987) reported similar results in that the 85 item test produced higher detection rates than a 30 item test. A limitation to studies by Rudner (1983) and Drasgow, Levine, and McLaughlin (1987) is that in comparing the effectiveness of the indices, test length may have been

confounded with test content. Therefore it was not known whether the reported differences between the two tests was attributable to test type, test length, or a combination of both. It can be concluded therefore that research is needed to investigate the effect of test length on appropriateness indices.

Appropriateness indices are based on examinee ability and item parameters, which can be estimated by various IRT models. Drasgow (1982) compared indices derived from the Rasch model and the 3PLM and although the results favoured the 3PLM, differences between the two models were small. In addition, Drasgow used non-standardized indices; the relationship between IRT model and detection rates of standardized indices has not been studied. Two researchers (Birenbaum, 1985; Parson, 1983) have used the 2PLM only. Researchers have not compared detection rates of indices for the 3PLM or 2PLM. It can be concluded that further research would be useful on the distributions and effectiveness of appropriateness indices under IRT models other than the 3PLM, especially the 2PLM.

Investigations of correlations among indices have shown that in general, ECI's and ECIZ's tend to be highly inter-correlated and that all of these are naturally negatively correlated with Lz. The unstandardized indices, Lo, Ln, and Lg, are less well correlated with the ECI's and ECIZ's. The correlations among U1, W1, U3, and W3 have been from .30 to .45. Similar correlation coefficients are reported using real data between each of Lz, Lo, and Ln with U1, W1, U3, and W3. There are no reported simulation

studies which have investigated the correlation among the three general categories of IRT based indices. It can be concluded that further research is needed on the correlations among indices.

In summary, the review of the research has identified several general concerns related to appropriateness measurement in both non-aberrant and aberrant response patterns. First, there is a need to examine the characteristics of specific appropriateness indices. Of the two general categories, IRT based and group-dependent, it appears that IRT based indices may be more promising because they can be standardized and are not dependent on group determined item difficulty. Three general types of IRT indices have been identified, maximum likelihood estimates, fit statistics, and extended caution indices. Specifically, there is a need to examine further Lz, W3, and ECIZ4 because each is representative of one general type of IRT index and each has been shown to be among the most effective indices in detecting unusual response patterns.

Second, with respect to Lz, W3, and ECIZ4 the following areas appear to need additional research: i) the characteristics of the distributions of the three appropriateness indices, ii) the effect of test length and IRT model on appropriateness indices, and iii) the effectiveness of the indices under combinations of test length, IRT model, type of aberrance, and level of aberrance. In the next section, two general research questions related to these three concerns are presented.

Research Questions

In this section, two general problems are presented in relation to the following three appropriateness indices: i) the standardized maximum likelihood function, Lz, ii) the fourth standardized extended caution index, ECIZ4, and iii) the 3PLM fit statistic, W3.

The first general problem concerns the characteristics of the distributions of the three appropriateness indices in non-aberrant response patterns. To examine this problem the following specific research question needs to be investigated: What is the effect of test length and IRT model on the characteristics of the distributions of three standardized appropriateness indices in known non-aberrant response patterns?

The second general problem concerns the effectiveness of the three appropriateness indices in detecting aberrant response patterns under different measurement conditions. To examine this problem, the following specific question needs to be investigated: To what extent are the three standardized indices effective in detecting aberrant response patterns under combinations of test length, IRT model, type of aberrance, and level of aberrance? In the next chapter procedures used to investigate these two general problems are presented and discussed.

CHAPTER 2

PROCEDURES

In this study two general questions related to three appropriateness indices, Lz, ECIZ4, and W3 are examined. The first general question is as follows: What is the effect of test length and IRT model choice on the distributions of three appropriateness indices: i) the standardized maximum likelihood function, Lz; ii) the fourth standardized extended caution index, ECIZ4; and iii) the three parameter fit statistic, W3? To investigate the first general problem, data were generated by computer in four combinations of two test lengths and two IRT models to produce non-aberrant response patterns.

The second general problem concerns the effectiveness of appropriateness indices in detecting aberrant response patterns. To examine this problem the following specific question was investigated: To what extent are the three indices effective in detecting aberrant response patterns under combinations of test length, IRT model, type of aberrance, and level of aberrance? To investigate this problem data were generated by computer in twenty-four combinations of two test lengths, two IRT models, two types of aberrance, and three levels of aberrance.

This chapter is organized as follows. First, the rationale for employing simulation techniques, and some limitations to such techniques, are discussed briefly. Second, procedures are

explained in which simulated data were generated for non-aberrant response vectors for combinations of two test lengths and two IRT models. Third, procedures used to analyze the characteristics of the distributions of the three indices in these four combinations of test length and IRT model are presented. Finally, procedures are explained in which aberrant response patterns were generated in twenty-four combinations of test length, IRT model, type of aberrance, and level of aberrance. The procedures used to determine the detection rates of the three appropriateness indices in these twenty-four combinations are also presented and discussed.

Advantages of Simulated Data

Appropriateness indices are used to analyze response vectors which are an examinee's responses, correct or incorrect, to a series of test items. Simulation procedures were employed to generate response vectors under a number of measurement conditions. Simulated data offer several advantages over data from real subjects. The first advantage is that with simulated data one can systematically examine the effectiveness of an appropriateness index. To do this it is first important to examine the distribution of appropriateness indices in non-aberrant response patterns. Using simulation techniques, response vectors can be generated from a known distribution of examinees for a test with known item parameters. Such response vectors can then be considered to be those of examinees who

respond appropriately with respect to their ability. Appropriateness indices are then computed for the generated non-aberrant response vectors and cut-off values set for false positive rates.

A second advantage to simulation techniques is that known types of aberrance and levels of aberrance can be produced in examinee response vectors. Spuriously high or spuriously low scores can be simulated by modifying response vectors to create specified levels for each of these two types of aberrance. Since the level of aberrance is known, it is possible to establish the extent to which the appropriateness indices can detect the aberrant response vectors at known false positive rates. Further, with simulation techniques it is also possible to generate tests of different lengths with specific item parameters.

A third advantage to simulation techniques is that different IRT models can be used to produce different conditions for types of aberrance and levels of aberrance. In this study it was possible to produce tests of different lengths for specified item parameters, to produce examinees of known abilities, to generate response vectors with different IRT models, and as noted above, to produce different types of aberrance and different levels of aberrance.

A final advantage to simulation procedures is that an experiment may be replicated many times over making it possible to examine the stability of the results over many samples.

Limitations of Simulated Data

Although simulation procedures enable systematic investigation of appropriateness indices, there are limitations to such procedures. First, the simulated item parameters are not data from real life because they are selected within given conditions and not estimated from an actual test; therefore, they may have limited generalization. Second, simulated response vectors may be affected by randomness moreso than real response vectors; presumably for real life examinees guessing rates may be different than in simulated conditions. Third, factors such as omit rates may affect the response pattern in real test situations and consequently the distribution of an appropriateness index (Drasgow, Levine, & Williams, 1985). Data for this study, for example, were generated with no omit rates. In spite of these limitations the opportunity to investigate known levels of aberrance justifies the use of simulated data.

In the following sections the procedures used to investigate appropriateness indices in both non-aberrant and aberrant conditions are presented.

Research Design for Examining the Distributions of Indices for Non-Aberrant Response Patterns

Non-aberrant response vectors are those which are expected for an examinee of a particular ability. In this study, non-aberrant response vectors were generated for four conditions based on combinations of two test lengths and two IRT models. To

examine the effects of test length on appropriateness indices, tests of two lengths, 40 items and 80 items, were simulated.

To examine the effects of the IRT model, response vectors were generated using both the 2PLM and the 3PLM. Consequently, data were generated according to a 2 X 2 (test length and IRT model) fully crossed design. The three appropriateness indices were then computed. The data generation and analysis procedure were replicated 50 times.

As described in Table 1, non-aberrant response vectors were generated in Steps 1 through 3 for simulated test items with known distributions of item parameters and examinee abilities. In Step 1 item parameters were generated using Datagen, a fortran computer program developed by Hambleton and Swaminathan (1977) and modified by Carlson (1985). In Datagen, appropriate IMSL subroutines are used to randomly generate item parameters from specified ranges and distribution.

The ranges of the item difficulties and item discriminations were similar to those suggested by Hambleton and Swaminathan (1985) for a general achievement test. For both the 2PLM and the 3PLM item difficulty parameters were selected from a range of -2.00 to +2.00 from a uniform distribution. Item discrimination parameters were selected from a range of .40 to 1.50 also from a uniform distribution for both the 3PLM and the 2PLM. For the 3PLM, the pseudo-chance or guessing parameters were allowed to

Table 1

Data Generation and Analysis for Non-Aberrant Response Patterns

Step 1. Item difficulty parameters were selected from a range of -2.00 to 2.00 from a uniform distributions for two lengths of tests (40 item and 80 item) using each of two models (2PLM and 3PLM). Item discrimination parameters were selected from a uniform distribution from .40 to 1.50; guessing parameters were selected from a range of .05 tO .20 for the 3PLM and set to 0 for the 2PLM.

Step 2. Four samples of abilities (N = 2000) from the normal (0,1) distribution were generated using Datagen (1985).

Step 3. With the generated item parameters (Step 1) and generated abilities (Step 2), Datagen (1985) was used to generate response vectors in four combinations of 2 test length and 2 IRT models using one sample (N = 2000) of abilities in each.

Step 4. Three appropriateness indices, Lz, ECIZ4, and W3, were computed for each examinee in each of the four combinations of test length by IRT model.

Step 5. The mean, standard deviation, skewness, and kurtosis were computed for each index in each of the four combinations of test length and IRT models. Correlations among the three indices were computed for each of the four conditions.

Step 6. The values for each index were computed at four false positive rates. To do this, they were identified at P_{99}, P_{95}, P_{90}, and P_{75} for ECIZ4 and W3 and at P_{01}, P_{05}, P_{10}, and P_{25} for Lz.

Step 7. Steps 1 to 6 were replicated a total of 50 times.

Step 8. The mean, standard deviation, minimum and maximum values over 50 replications were calculated for the mean, standard deviation, skewness, and kurtosis. This was done for each index in each of the four conditions. To examine the effects of test length and model choice, a 2 X 2 MANOVA was also performed on the means of each index from the 50 replications.

Step 9. For each percentile (false positive rate) the mean and standard deviation over 50 replications were computed for each index. To examine the effects of test length and IRT model, a 2 X 2 MANOVA was performed using test length and IRT model as independent variables and the values of each index at each false positive rate as dependent variables.

range from .05 to .20 in a uniform distribution. For the 2PLM this parameter was set to 0. The ranges and distributions of the item parameters were the same for each of the 40 item and 80 item tests with different seed numbers used to generate the parameters for each length of test. The same seed number was used for the 40 item test for both the 2PLM and the 3PLM and another seed for the 80 item test for both the 2PLM and the 3PLM.

In Step 2, simulated examinees were generated from a normal (0,1) distribution of abilities with the same seed used for each combination of test length and IRT model. The range of these abilities was from approximately -3.50 to 3.50 with 2000 generated in each of the four conditions.

In Step 3, Datagen (1985) was used to generate a response vector for each examinee in each of the four conditions using the generated item parameters and abilities. In summary, data were generated for 2000 examinees for each test length by IRT model. It is important to emphasize that in these procedures each examinee is considered to be responding appropriately with respect to ability.

In Step 4, the three standardized appropriateness indices Lz, ECIZ4, and W3 were computed for each of the 2000 examinees in each of the four conditions using a computer program developed by Drasgow (1985). It should be noted that under the 2PLM W3 should be more properly labelled W2; however, for the sake of consistency W3 is used for the index in both the 3PLM and the 2PLM. Generated item parameters, generated examinee abilities,

and response vectors were used to compute each of the three indices. The generated response vectors are treated as observed responses for each examinee. Using the generated item parameters and abilities, it is possible to estimate an expected pattern of responses.

In Step 5, the mean, standard deviation, skewness, and kurtosis were computed for each index in each of the four conditions of test length by IRT model. Also in Step 5 the Pearson product-moment correlations among the indices were computed in each of the four conditions.

In Step 6, the value of each index at each of four percentiles was computed. Since for Lz, smaller values indicate more inappropriate response patterns, the following percentiles were used: P_{01}, P_{05}, P_{10}, and P_{25}. For ECIZ4 and W3, because larger values indicate more inappropriate patterns, the percentiles used were P_{99}, P_{95}, P_{90}, and P_{75}. These percentiles were selected since they represented commonly used false positive rates.

The first six steps were replicated a total of fifty times in Step 7. Each replication consisted of generated data on 2000 subjects on each of a 40 item and 80 item test according to each of a 3PLM or a 2PLM. New seed numbers were used to generate all item parameters and abilities for each replication. The ranges and distributions for the item parameters and abilities were the same for each replication.

In Step 8, the mean, standard deviation, minimum and maximum

values of each index were computed over the 50 replications. These four statistics were computed for the mean, standard deviation, skewness, and kurtosis for each index in each of the combinations of test length and IRT model. This information was used to examine the stability of the shape and location of the distributions of the indices over samples. Also in Step 8, a 2 X 2 MANOVA was performed using the mean values for each index over the fifty replications as the dependent variables and test length and IRT model as independent variables. Finally, the mean correlations among indices were computed over the 50 replications for each of the four combinations.

In Step 9, the values of each index at each of the four percentiles were used to examine other aspects of stability of the distributions of the indices over the 50 replications. The mean and standard deviation were computed for each index in each of the four conditions of test length and IRT model at each of the four percentiles over the 50 replications. To further examine the effects of test length and model choice, a 2 X 2 MANOVA was performed using the four percentiles of each index as dependent variables. For each index and each percentile the mean value over the 50 replications was used to determine a cut-off for false positive rate.

Research Design for Determining the Effectiveness of Indices for Aberrant Response Patterns

This section includes a description of the procedures used

to generate aberrant response patterns for twenty-four combinations of test length, IRT model, type of aberrance, and level of aberrance. A summary of the procedures used in Steps 10 to 14 is shown in Table 2. Researchers have reported two types of aberrant response patterns, spuriously high and spuriously low, each of which may exist at many levels. Levels of aberrance may vary in different testing situations; however, Drasgow, Levine, and McLaughlin (1987) suggested that 15% and 30% aberrance represented low and moderate levels of aberrance respectively. Other researchers have used 5%, 10%, 20%, and 40% aberrance. For this study, three levels of aberrance were used, 10%, 15%, and 30%. These three levels can be considered realistic for practical purposes.

Item parameters and examinee abilities were used to generate modified response patterns for each of 24 combinations of two test lengths (40 item and 80 item), two IRT models (2PLM and 3PLM), two types of aberrance (spuriously high and spuriously low), and three levels of aberrance (10%, 15%, and 30%).

In Step 10, item parameters were generated from uniform distributions as follows. For the 3PLM the difficulty parameters were selected from a range of -2.00 to 2.00; the discrimination parameters ranged from .40 to 1.50, and the guessing parameters ranged from .05 to .20. For the 2PLM the same ranges and distributions were used except that the guessing parameter was specified to be 0. To ensure that the 80 item test was an extension of the 40 item test, the same seed numbers were used

Table 2

Data Generation and Analysis for Aberrant Response Patterns

Step 10. Using Datagen (1985) item parameters were generated from specified ranges from uniform distributions for two test lengths, 40 item and 80 item, and for two IRT models, the 3PLM and the 2PLM. Examinee abilities (N = 4000) were generated from a normal (0, 1) distribution for each of four combinations of test length and IRT model.

Step 11. Datagen (1985) was used to produce response vectors for the four combinations of test length and IRT model. Six copies of each of the four combinations were then produced to provide a total of 24 sets of response patterns of 4000 examinees each. These sets of response patterns were subsequently modified to create 24 combinations of aberrant response pattterns.

Step 12. Aberrant response patterns were created by modifying 10, 15, and 30 per cent of the items for each examinee to produce the 24 combinations of response patterns.

Step 13. The ability of each examinee was re-estimated and the three appropriateness indices computed for each of the 24 combinations. Means and standard deviations for the indices were also computed.

Step 14. For each index the mean values of each index at each of the four percentiles (Step 9) was used as a cut-off value with the percentile considered as a false positive rate. Detection rates for each false positive rate were computed for each index in each of the twenty-four combinations.

---

for each of the two tests for each of the 3PLM and the 2PLM. Using the same seed examinee abilities (N = 4000) were generated from a normal (0, 1) distribution for each of the four combinations.

In Step 11 Datagen was used to generate response vectors, for each of the 4000 examinees in the four combinations of test length and IRT model using the generated item parameters and abilities from Step 10. Six copies of each of the four

combinations of response patterns were produced; thus 24 sets of response patterns of 4000 examinees each were available for modification to create aberrant patterns

In Step 12 the 24 sets of response patterns were modified to produce 24 combinations of two test lengths, two IRT models, two types of aberrance and three levels of aberrance. Three levels of aberrance, 10%, 15%, and 30%, were used for each of the spuriously high and spuriously low conditions. Spuriously high scores were produced by randomly sampling a specified percentage of incorrect items 10%, 15%, or 30%, without replacement for each examinee. The sampled item was then re-scored from incorrect(0) to correct(1). For example, in the case of 10 percent spuriously high aberrance on the 40 item test four items would be re-scored for each examinee. Spuriously low scores were obtained by sampling 10%, 15%, or 30% of the correct items from each examinee and re-scoring the item from correct to incorrect. As a check that response vectors were modified, mean scores of modified response vectors were computed. These scores were compared with the unmodified response vectors so it was possible to confirm that the response vector had been modified spuriously high or spuriously low as designed.

In Step 13 the three appropriateness indices were computed for each of the 4000 examinees in each of the twenty-four conditions using a fortran program (Drasgow, 1985). Because modifying the response vector can be considered to have the effect of changing the original ability of the examinee, the

ability of each examinee was re-estimated within the computer program prior to computing the index. Also in Step 13 the mean and standard deviation were computed for each index for each of the twenty-four combinations.

The effectiveness of an appropriateness index can be measured by the extent to which it is able to detect aberrant response patterns. A technique employing the receiver operating curve was first introduced by Levine and Rubin (1979) to examine detection rates when samples of examinees of both aberrant and non-aberrant response patterns are available. Subsequently, Parsons (1983), Drasgow and Levine (1982), Drasgow, Levine, and Williams (1985) and Drasgow, Levine, & McLaughlin (1987) have used similar procedures. In this study a slightly different procedure was used. The cut-off scores had been identified at four percentiles for each index (see Step 9). The four percentiles were used to establish false positive rates of .01, .05, .10, and .25.

In Step 14 detection rates were computed for each index in each of the 24 combinations of test length, IRT model, type of aberrance, and level of aberrance. The detection rates of the three indices were then compared under each of the twenty-four combinations.

Because modifying response patterns may have the effect of changing the ability of a simulated examinee, an additional analysis was conducted to compare detection rates using true (generated) ability and estimated (modified) ability. To examine

this effect, six categories of true ability were established with cut-off points at -1.50, -.60, 0, .60, and 1.50. Six categories of ability estimated from response patterns were also established for each level of aberrance for each index. Since modified responses affect distributions of ability, different cut-offs were established for each type of aberrance and level of aberrance based on what was considered to be a realistic distribution. The detection rate was then computed for each index at each false positive rate in each category for both the original abilities and the modified abilities in each of the 24 combinations. Thus it was possible to compare detection rates in each ability category.

Summary

In summary, data for this study were generated by computer for both simulated non-aberrant response patterns and simulated aberrant response patterns. For non-aberrant response patterns, data were produced in combinations of two test lengths and two IRT models. For aberrant response patterns, 24 conditions were established for combinations of two test lengths, two IRT models, two types of aberrance, and three levels of aberrance. The characteristics of the distributions of three appropriateness indices in both non-aberrant and aberrant response patterns were investigated using several procedures. The effects of test length and IRT model on the distribution of indices were examined for non-aberrant response patterns. The detection rates for each

index in each of the twenty-four combinations were computed for aberrant response patterns. In the next chapter the results of the study are presented and discussed.

CHAPTER 3

RESULTS AND DISCUSSION

In this chapter the results of the study are presented and discussed. In the first section, the effects of test length and IRT model on the distribution of three appropriateness indices, Lz, ECIZ4, and W3, in non-aberrant response patterns are presented and discussed. In the second section, the distribution and effectiveness of the indices in aberrant response patterns are presented and discussed. Finally, a summary of the results and discussion are presented.

Appropriateness Indices for Non-Aberrant Response Patterns

Researchers such as Tatsuoka and Tatsuoka (1982a,1982b) and Drasgow, Levine, and Williams (1985) have emphasized that to be of value an appropriateness index must not be dependent upon examinee ability. As a preliminary check prior to presenting the results, the correlations between examinee ability and the indices were computed for a large sample (N=10,000) for each of the four combinations of test length and IRT model. The results, which are presented in Table 3, confirm that there were no significant correlations between any of three indices and examinee ability. This certainly suggests that, for non-aberrant response patterns, there is no significant linear relationship between examinee ability and the three appropriateness indices.

The correlations between absolute values for ability and the indices were also examined and found to be very similar to those in Table 3. This was done because of earlier findings that extreme ability values were associated with high index values for non-standardized indices.

Table 3

Correlations Between Ability and Appropriateness Indices

	θ/Lz		θ/ECIZ4		θ/W3	
	r	(p)	r	(p)	r	(p)
40 item/3PLM	-.011	(.253)	-.009	(.392)	.013	(.193)
40 item/2PLM	-.017	(.084)	-.010	(.313)	.016	(.105)
80 item/3PLM	-.009	(.351)	-.004	(.682)	.013	(.178)
80 item/2PLM	-.006	(.552)	.008	(.447)	.018	(.078)

Shape and Location of the Distribution of Indices for Non-Aberrant Response Patterns

Tables 4, 5, and 6 present the summary statistics for each of Lz, ECIZ4, and W3, respectively, for each of the four combinations of test length and IRT model over fifty replications. The mean, standard deviation, range, skewness, and kurtosis for each index are presented. With respect to the means results were produced which were very close to expected values for each index over the four combinations of test length and IRT

Table 4

Descriptive Statistics for Lz Over Fifty Replications

Statistic		40 item/2PLM	40 item/3PLM	80 item/2PLM	80 item/3PLM
Mean	M	-.007	-.009	-.004	-.005
	SD	.022	.021	.022	.022
	Min	-.059	-.045	-.061	-.041
	Max	.033	.034	.039	.052
Standard Deviation	M	1.003	1.003	.999	.999
	SD	.016	.015	.016	.017
	Min	.980	.975	.968	.965
	Max	1.045	1.041	1.041	1.051
Skewness	M	-.621	-.513	-.432	-.355
	SD	.081	.069	.063	.063
	Min	-.851	-.630	-.590	-.586
	Max	-.452	-.324	-.282	-.247
Kurtosis	M	.509	.340	.222	.139
	SD	.297	.196	.223	.209
	Min	.056	-.094	-.155	-.213
	Max	1.442	.776	.992	1.092

Table 5

Descriptive Statistics for ECIZ4 Over Fifty Replications

Statistic		40 item/2PLM	40 item/3PLM	80 item/2PLM	80 item/3PLM
Mean	M	.004	.004	.003	.004
	SD	.020	.018	.024	.023
	Min	-.031	-.028	-.043	-.041
	Max	.050	.043	.072	.077
Standard Deviation	M	.999	.999	1.002	1.001
	SD	.015	.015	.014	.015
	Min	.949	.962	.963	.957
	Max	1.026	1.032	1.029	1.029
Skewness	M	.277	.232	.205	.177
	SD	.065	.066	.048	.051
	Min	.121	.114	.118	.069
	Max	.455	.378	.326	.256
Kurtosis	M	.014	-.022	-.008	-.036
	SD	.160	.115	.114	.113
	Min	-.215	-.248	-.199	-.336
	Max	.748	.437	.319	.176

Table 6

Descriptive Statistics for W3 Over Fifty Replications

Statistic		40 item/2PLM	40 item/3PLM	80 item/2PLM	80 item/3PLM
Mean	M	1.002	1.002	1.000	1.000
	SD	.005	.004	.003	.003
	Min	.993	.993	.994	.992
	Max	1.013	1.012	1.008	1.006
Standard Deviation	M	.232	.205	.162	.144
	SD	.011	.011	.007	.007
	Min	.214	.181	.149	.127
	Max	.256	.229	.186	.160
Skewness	M	.624	.580	.413	.328
	SD	.250	.306	.241	.228
	Min	.210	.081	-.010	-.092
	Max	1.528	1.711	1.079	1.084
Kurtosis	M	2.599	2.990	2.211	1.957
	SD	2.342	3.212	1.998	1.691
	Min	.539	.295	-.053	.055
	Max	12.142	16.633	9.311	9.290

model. The means for Lz and ECIZ4 are very close to 0 while those for W3 are approximately 1.0. The variability of the means over replications was slightly larger for Lz and ECIZ4 than for W3.

The mean standard deviation of both Lz and ECIZ4 over replications was approximately 1, with each showing similar variability over the 50 replications. The mean standard deviation for W3 varied from .144 to .232 which indicated that W3 showed more variability than Lz or ECIZ4 over the four combinations of test length and IRT model. This suggests that test length and IRT model may affect the variability of the distribution of W3 but not that of Lz or ECIZ4. Within combinations of test length and model choice the standard deviation varied less for W3 than for Lz or ECIZ4.

The mean skewness of the distributions showed more variability than did the means for each index for each of the four conditions over the fifty replications. For Lz, mean skewness ranged from -.355 to -.621 for combinations of test length and IRT model. For ECIZ4, skewness ranged from .177 to .277 and for W3 the range was from .328 to .624 for the four combinations. Of the three indices, ECIZ4 exhibited the smallest amount of skewness. For each index the 80 item test produced smaller skewness values than the 40 item test and the 3PLM produced smaller values than the 2PLM. Further, for Lz and ECIZ4 the standard deviation of the skewness indices was considerably smaller than for W3. These results showed that with respect to

skewness, ECIZ4 deviated least from normality. W3 and Lz produced similar results for skewness. It is interesting to note that over the 50 replications for Lz each skewness value was negative and for ECIZ4 each value was positive. This suggested, as expected, that the tail of the distribution showing skewness was where extreme values indicate aberrant response patterns.

An examination of kurtosis over replications revealed that ECIZ4 produced much less kurtosis than either Lz or W3 for the four combinations of test length and IRT model. For Lz and W3 the value for the 40 item test was greater than for the 80 item test. For Lz and ECIZ4, kurtosis was greater for the 2PLM than for the 3PLM. In the case of W3 the values were quite large indicating a quite peaked distribution for that index. The standard deviations and ranges of the kurtosis indices indicated that ECIZ4 showed the least variability and W3 extremely large variability. In fact, the range of kurtosis indices for W3 was very large with maximum values more than 9.0 for each of the four combinations of test length and IRT model. This too suggested variability of the index over replications. With respect to kurtosis, the results showed that ECIZ4 deviated least from normality and Lz the next least. The distribution for W3 appeared to be non-normal.

In summary, the means and standard deviations of Lz and ECIZ4 were similar for each of the four combinations of test length and IRT model over 50 replications. For W3, the means were similar over the four combinations but the standard

deviations showed some variability over replications. All indices showed some skewness and kurtosis with W3 showing a much higher mean value and range for kurtosis in each of the four combinations than either Lz or ECIZ4. Of the three indices, ECIZ4 showed the least skewness and kurtosis and the least variability over replications; in fact, kurtosis did not seem to be a problem. ECIZ4 demonstrated the most stable distribution with respect to shape. For all indices, skewness and kurtosis seemed to be somewhat related to test length and IRT model. The skewness and kurtosis for W3 over replications indicated that the distribution of the index is non-normal and that the distribution is the most unstable of the three indices over replications. The distribution of Lz also showed more skewness and kurtosis than ECIZ4. The size and variability of skewness and kurtosis indices for Lz might lead one to question whether the index produces a near normal distribution as reported by Drasgow, Levine, and Williams (1985). Although it may be concluded that in non-aberrant examinee responses the mean values of indices demonstrate stability over samples, this is not true for the skewness and kurtosis of the indices as discussed above, or for the standard deviation of W3.

Effect of Test Length and IRT Model

A multivariate analysis of variance was conducted using test length and IRT model as the independent variables and the mean over replications of each index as dependent variable. The purpose of this analysis was to determine if test length, IRT

model, or their interaction had an effect on values of the mean of the means over 50 replications.

A summary of the results of the multivariate analysis of variance of the means as a function of test length and model choice are presented in Table 7. For the multivariate test,

Table 7

Tests of the Effect of Test Length and IRT Model on the Means of Three Appropriateness Indices

Multivariate Analysis of Variance		
Effect	Multivariate	(df 3, 194)
	F	p
Test Length	.38	.76
IRT Model	2.67	.049
Length X Model	.02	.99

Post Hoc Univariate Analysis of Variance						
	Univariate			(df 1, 196)		
	Lz		ECIZ4		W3	
	F	p	F	p	F	p
IRT Model	1.03	NS	.13	NS	3.68	NS

results showed that over 50 replications only the IRT model effect was significant ($p < .05$). For the follow-up univariate analysis of variance there were no significant effects. This was interpreted to mean that over the fifty replications (N = 2000), the mean of the mean does not vary significantly with respect to test length, IRT model, or their interaction.

Determining False Positive Rates

The effects of test length and IRT model on the tails of the distribution of the indices in the non-aberrant response patterns were also examined. The purpose of this procedure was to establish cut-offs for false positive error rates. The value of each index was obtained at four percentiles for each replication and the mean computed over the 50 replications. For Lz, the following percentiles were used P_{01}, P_{05}, P_{10}, and P_{25}; for ECIZ4 and W3, P_{99}, P_{95}, P_{90}, and P_{75} were used. Subsequently, the mean value of each index at each percentile was considered as a cut-off score for false positive rate to use in detecting aberrant response patterns. The mean value over fifty replications for each index at each percentile in each of the four conditions is presented in Tables 8, 9, and 10.

Of the three indices, the values of Lz appeared to deviate the most from the mean over the 50 replications for each of the combinations of test length and IRT model at each of the four percentiles. Although the distributions for Lz and ECIZ4 both had a mean of 0 and standard deviation of 1, the values of Lz at three of the four percentiles deviated more from the mean than

Table 8

Mean and Standard Deviation of Lz at Selected Percentiles Over Fifty Replications

		Percentile							
		P_{01}		P_{05}		P_{10}		P_{25}	
Test Length	IRT Model	M	SD	M	SD	M	SD	M	SD
40 item	3PLM	-2.724	.120	-1.799	.053	-1.347	.041	-.632	.034
40 item	2PLM	-2.791	.139	-1.826	.062	-1.348	.043	-.620	.030
80 item	3PLM	-2.593	.109	-1.750	.054	-1.328	.041	-.646	.029
80 item	2PLM	-2.652	.120	-1.771	.054	-1.328	.051	-.639	.033
Marginals									
Length	40	-2.758	.137	-1.812	.059	-1.348	.042	-.626	.032
	80	-2.623	.118	-1.760	.055	-1.328	.046	-.642	.030
IRT Model									
	3PLM	-2.659	.132	-1.775	.059	-1.337	.042	-.639	.032
	2PLM	-2.721	.147	-1.798	.064	-1.338	.048	-.629	.033

did those of ECIZ4. Further, the standard deviation of Lz was somewhat greater than that of ECIZ4 at each percentile indicating more variability over replications in each of the four conditions. The values of W3 at each percentile appeared to deviate less from the mean and to show less variability than either Lz or ECIZ4. This might be expected given the low

Table 9

Mean and Standard Deviation of ECIZ4 at Selected Percentiles Over Fifty Replications

		Percentile							
		P_{99}		P_{95}		P_{90}		P_{75}	
Test Length	IRT Model	M	SD	M	SD	M	SD	M	SD
40 item	3PLM	2.466	.089	1.715	.047	1.322	.042	.661	.028
40 item	2PLM	2.522	.098	1.715	.046	1.315	.043	.661	.029
80 item	3PLM	2.454	.095	1.698	.046	1.309	.034	.673	.025
80 item	2PLM	2.492	.091	1.714	.047	1.312	.040	.669	.032
Marginals									
Length	40	2.495	.097	1.715	.046	1.318	.043	.661	.029
	80	2.473	.094	1.706	.047	1.311	.037	.671	.029
IRT Model									
	3PLM	2.460	.092	1.707	.047	1.316	.038	.667	.027
	2PLM	2.508	.095	1.714	.046	1.314	.042	.665	.031

standard deviation of that index over replications.

The marginal mean values were used to examine mean index values for different test lengths across the four percentiles. For each index the 40 item test produced more extreme values than the 80 item test over percentiles except in the case of P_{25} for Lz and P_{75} for ECIZ4 where the 80 item test produced a more

Table 10

Mean and Standard Deviation of W3 at Selected Percentiles Over Fifty Replications

		Percentile							
		P_{99}		P_{95}		P_{90}		P_{75}	
Test Length	IRT Model	M	SD	M	SD	M	SD	M	SD
40 item	3PLM	1.555	.047	1.347	.019	1.259	.013	1.123	.078
40 item	2PLM	1.629	.038	1.398	.022	1.294	.015	1.139	.009
80 item	3PLM	1.372	.029	1.240	.012	1.181	.009	1.089	.005
80 item	2PLM	1.423	.024	1.273	.013	1.205	.010	1.010	.006
Marginals									
Length	40	1.592	.057	1.373	.033	1.277	.022	1.132	.012
	80	1.396	.037	1.2562	.021	1.193	.015	1.094	.007
IRT Model									
	3PLM	1.464	.100	1.294	.056	1.220	.041	1.106	.019
	2PLM	1.526	.108	1.335	.066	1.249	.047	1.120	.021

extreme value. Differences in marginal means for test length were smallest for ECIZ4 and largest for W3. An examination of the standard deviation of the marginal means showed that W3 produced the least, and Lz the most variation at each percentile.

The marginal means for the two IRT models indicated that the 2PLM produced more extreme values than the 3PLM for W3 at each

percentile. For Lz, the 2PLM produced more extreme values at P_{01}, P_{05}, and P_{10}. For ECIZ4 the 2PLM was greater for P_{99} and P_{95} but for the 3PLM values were similar for P_{90} and P_{75}. For each index the 80 item 3PLM combination had the lowest absolute value except for ECIZ4 at P_{75}. As with test length, ECIZ4 seemed least affected by IRT model.

In Table 11, results are summarized for a multivariate analysis of variance which was performed to examine the effect of test length and IRT model for the three indices at the four false positive rates. Using .05 level of significance, the multivariate test showed that each of the independent variables and the interaction produced a significant effect. In the follow-up univariate analysis of variance for Lz the differences were significant for test length at all percentiles and for IRT model at the .01, .05, and .25 false positive rates. There were no significant interaction effects. For ECIZ4, only IRT model was significant at .01 false positive rate and test length at .75 false positive rate. There were also no significant interaction effects. For W3, all main effects and interactions were significant at all false positive rates.

Of the three indices, the distribution of ECIZ4 was least affected by test length and IRT model, Lz, the second least and W3 most affected by test length, IRT model and the interaction of test length and IRT model. Again, these results suggest that the distribution of ECIZ4 was more stable and was least affected by test length and IRT model and W3 was most affected.

Table 11

Effects of Test Length and IRT Model on Three Appropriateness Indices at Various False Positive Rates

Multivariate Analysis of Variance (df 12,188)

Test Length		IRT Model		Interaction	
F	p	F	p	F	p
251.77	.00	33.38	.00	1.85	.04

Univariate Analyses of Variance (df 1,196)

Index	False Positive Rate	F	p	F	p	F	p
Lz	.01	60.63	.00	13.03	.00	.04	.84
	.05	43.76	.00	8.76	.00	.15	.70
	.10	10.14	.00	.02	.88	.01	.93
	.25	13.17	.00	4.95	.03	.32	.57
ECIZ4	.99	2.60	.11	12.80	.00	.50	.48
	.95	1.88	.17	1.41	.23	1.52	.22
	.90	1.83	.18	.14	.70	.70	.40
	.75	5.62	.02	.36	.55	.21	.65
W3	.99	1464.17	.00	150.06	.00	4.85	.03
	.95	2325.92	.00	299.94	.00	13.93	.00
	.90	2474.95	.00	303.61	.00	9.61	.00
	.75	1445.67	.00	184.94	.00	7.24	.01

Correlations Among Indices

The final analysis of the three indices in non-aberrant response patterns involved the correlations among the three indices. Table 12 presents the mean correlations among the indices over the fifty replications. For each correlation the 80 item test produced somewhat higher correlations than the 40 item test and the 3PLM produced higher correlations than the 2PLM.

Table 12

Mean Intercorrelations Among Indices Over Fifty Replications

	Lz/ECIZ4	Lz/W3	ECIZ4/W3
3PLM/40 Item	-.674	-.949	.581
2PLM/40 Item	-.652	-.944	.571
3PLM/80 Item	-.682	-.951	.594
2PLM/80 Item	-.661	-.946	.582

For Lz and ECIZ4 the correlations ranged from -.652 to -.682 over the four combinations of test length and IRT model. Lz and W3 produced the highest correlations with a range of -.944 to -.951 over the four combinations of test length and IRT model. ECIZ4 and W3 produced the lowest correlations, which ranged from .571 to .594, across the four combinations.

In summary, the analysis of the distributions of the indices in non-aberrant response patterns has provided information on the effects of test length and IRT model over 50 replications. The

means and standard deviations of the indices were found to be stable over replications; however, variability in the skewness and kurtosis indices suggested that those indices for Lz and W3 are not necessarily stable over replications. Further, the 3PLM produced less skewness and kurtosis than did the 2PLM for each index. As well, the 80 item test produced less skewness and kurtosis than did the 40 item test.

W3 seemed to be least stable over replications and showed extreme variability for kurtosis. For both skewness and kurtosis, ECIZ4 showed least departure from normality. Lz and W3 had similar values for skewness; however, for kurtosis W3 produced very extreme values. Thus it appears that ECIZ4 produced the most nearly normal distribution, Lz next and W3 the least normal distribution.

The distributions were also examined at four percentiles and the value of each index at each percentile was treated as a cut-off score. Analysis of the indices at four cut-off scores indicated that test length and the IRT model affected the distributions of each index differently with W3 being most affected. In the following section the results of the study with respect to the distribution of the indices in non-aberrant response patterns are discussed.

Discussion

Analysis of the correlations between ability and the indices found that there was no linear relationship between the indices and examinee ability. This supports the assumption that the

indices are independent of ability and therefore independent of total score. This is what is expected because the indices are considered to be standardized and therefore there would likely be little relationship between each index and examinee ability. This is important since the values of the indices then have the same meaning at any ability level. Also, extreme values of ability did not seem to be related to size of any of the three indices. Previously Birenbaum (1985) and Drasgow, Levine, and McLaughlin (1987) found that standardized indices including ECIZ4 and Lz were not curvilinearly related to total test score. Also, Drasgow, Levine, and McLaughlin (1987) found W3 to be well standardized; therefore one might expect that W3 is not related to examinee ability. These results support the conclusion that the indices are not related to total test score.

It is desirable that an appropriateness index demonstrate stability for various test conditions, such as test length and IRT model. Further, skewness and kurtosis should be known for distributions of indices in non-aberrant response patterns. Such information is helpful if one is to establish meaningful detection rates for aberrant response patterns. Analysis of the means of the three indices over the 50 replications revealed little variation with respect to test length or IRT model. The means of each of the 50 distributions of Lz and ECIZ4 are approximately zero with a standard deviation near 1.0. These values can be considered close to the expected value insofar as Tatsuoka and Tatsuoka (1982) found the distributions of ECIZ4 to

be near normal (0,1) and Drasgow, Levine, and Williams (1985) found Lz to be approximately normal (0,1). For W3 the mean was approximately 1.0 which is the expected mean for a weighted mean square fit statistic (Grosse & Wright, 1988). The standard deviation for W3 ranged from .144 to .232 depending on test length and IRT model.

Comparison of means and standard deviations produced in this study with those reported by Drasgow, Levine, and McLaughlin (1987) revealed some similarities. Those researchers reported results of empirical distributions (N = 1000) for the 85 item SAT-V and found the mean and standard deviation for Lz to be .09 and .97 respectively. For ECIZ4 the mean and standard deviation was -.14 and .86 respectively and for W3 the values were .99 and .12. Overall, the results for this study are similar to those reported by Drasgow, Levine, and McLaughlin (1987) with respect to the means and standard deviations. However, the small differences in the means and standard deviations of the indices between this study and the earlier research may be the result of a number of factors. The earlier researchers used a smaller sample and only one replication. Also, examinee ability was re-estimated prior to computing the indices in the earlier research. Finally, this research used generated item parameters with uniform distributions rather than those estimated from an actual test as used by Drasgow, Levine, and McLaughlin (1987).

Other comparisons may be made with results of researchers who used data from actual persons. Birenbaum (1985) used a 20

item test and reported the mean and standard deviation for Lz to be -.32 and 1.61 and for ECIZ4 .38 and 1.196 respectively. W3 was not used in that study. Harnisch and Tatsuoka (1983) using a 68 item test, reported the mean and standard deviation for Lz to be .25 and 1.14 and for ECIZ4 as -.20 and 1.09. These results may not provide useful comparison for the results of this study since in both cases samples were much smaller and in the case of Birenbaum, the test was much shorter with some known aberrancy introduced by adding thirty simulated subjects to the data. Thus the differences in means and standard deviations can be expected since the data were not truly non-aberrant.

The mean standard deviation of each index over the fifty replications indicated that for Lz and ECIZ4 the distributions showed no variation over test length and IRT model. However, for W3 both test length and IRT model had a small effect on the mean standard deviation of the index over the 50 replications. That is to say the distribution for the 40 item test showed more variability than the 80 item test and the distribution of the 2PLM showed more variability than the 3PLM.

An analysis of variance of the means produced no significant effects by IRT model, test length, or interaction. This suggests that the means of the three indices demonstrate stability over four combinations of test length and IRT model.

Skewness and kurtosis estimates produced less consistent results over replications than did the means and standard deviations. Skewness estimates were smallest for ECIZ4. For

kurtosis, values for ECIZ4 are near 0 for the four combinations of test length and IRT model. The smaller standard deviation of both skewness and kurtosis over 50 replications also suggests that ECIZ4 is the least affected by test length and IRT model. This can be interpreted to mean that ECIZ4 produces a distribution closer to normal than either Lz or W3.

Lz appeared to be the second least affected by test length and IRT model with respect to skewness and kurtosis. Both skewness and kurtosis are larger and more variable than for ECIZ4. For W3, the large means and extreme variability for the kurtosis values indicated a peaked distribution which deviated from normality for all four combinations of test length and IRT model. The skewness indices were similar to those of Lz though more variable. It would appear that, compared to W3, Lz produced a more stable distribution which was closer to normal and less affected by test length and IRT model.

Analysis of the cut-off scores of the indices provided information on the effect of test length and IRT model, as well as information on the stability of the distributions over replications. The four percentiles were considered as the following false positive rates; .01, .05, .10, and .25. For Lz and ECIZ4 the cut-off values over replications showed a greater deviation from the mean than did those of W3. This may be related to the fact that the distributions of W3 were both less variable and leptokurtic. As expected, standard deviations of each index were greatest at the .01 false positive rate, which

indicated that the values are least stable in the extreme tails of the distributions. W3 had the smallest standard deviation at each false positive rate. The range of values over the four false positive rates was greatest for Lz, slightly less for ECIZ4, and least for W3. For Lz, the range was from -2.79 to -.62, for ECIZ4, the range was from .66 to 2.52, and for W3, 1.08 to 1.63.

The actual values of the cut-off scores for appropriateness measurement have not typically been reported in the research. It is therefore difficult to compare the cut-offs used in this study with those of other researchers. However, Drasgow, Levine, and Williams (1985), using the SAT-V, reported a cut-off of -1.64 for P_{05} for Lz. Drasgow and Guertler (1987) reported a value of -1.45 at P_{05} and -2.21 at P_{01} for Lz. In this study, the mean value for the 80 item test under the 3PLM was -1.75 at P_{05} and -2.59 at P_{01}. Differences between the cut-off values found in this study from those reported in the earlier study could be due to a number of reasons. First, the earlier researchers used re-estimated ability rather than generated ability. Such ability estimates take into account the modified item responses which has the effect of producing less extreme indices. Second, those researchers used estimated rather than generated item parameters, a slightly longer test, a smaller sample, and one replication whereas this study used a large sample replicated 50 times which could affect the estimates of the cut-off values. No comparable research studies using the 2PLM have been reported; it is not

possible therefore to compare the cut-off scores found here with any others.

The analysis of variance of the values of the indices at four percentiles revealed that W3 was most affected by test length and IRT model, Lz next most, and ECIZ4 the least affected. This is of interest because ECIZ4 was also the index which seemed to produce a distribution closest to normal, based on the mean, standard deviation, skewness, and kurtosis over 50 replications. This suggests that ECIZ4 is a more predictable index and is least affected by varying test conditions. In practical applications of the indices such information would be important.

With respect to the correlations among indices, the high correlations between Lz and W3 over the fifty replications indicated that the two indices may be similar in the extent to which they identify unusual response patterns. Because of the smaller amount of variability in index values for W3, it would seem to be more subject to effects of factors such as test length and IRT model. Given that these factors affect results significantly and the narrow range of index values, it is somewhat surprising that it correlates so highly with Lz. The correlations between Lz and W3 may result because the two indices are both based on differences between the probability of a particular response pattern and an actual response pattern for a particular ability. Lz is based on the likelihood of a particular response pattern; while W3 is based on the difference between the expected and observed response for each item.

Therefore, although the actual values of the Lz and W3 may differ, the correlations between the two indices may be high since they are based on similar principles. On the other hand, ECIZ4 is based on the ratio of two covariances related to observed and expected response patterns and is influenced by average probability of the group response whereas Lz and W3 are not so influenced. This may explain why ECIZ4 is not highly correlated with Lz and W3. The correlations under the 3PLM were higher than those under the 2PLM. Since the distributions of the indices tended to show less skewness and kurtosis under the 3PLM, this might explain the small differences. Similarly, the 80 item test tended to produce somewhat more stable distributions than the 40 item test, which might explain the slightly higher correlations for the 80 item test.

In summary, analysis of the three appropriateness indices in non-aberrant response patterns provided information about the distributions of the indices under four combinations of test length and IRT model. Analysis of cut-off scores indicated that ECIZ4 is least affected by test length and IRT model, although IRT model did have an effect on the cut-off scores of the index at .01 false positive rates. ECIZ4 produced the most stable distributions over replications and was least affected by test length and IRT model. The distribution of ECIZ4 appears to be approximately normal (0,1). For Lz, test length had an effect for all cut-off scores, IRT model had an effect for three of the four cut-off scores. Lz also produced an approximately normal

(0,1) distribution, although it departed more from normality and was somewhat less stable over replications than ECIZ4. For skewness and kurtosis, W3 was the least stable of the indices over replications and was most affected by test length and IRT model. The index produced a peaked distribution with a mean kurtosis greater than 2.0 for three of the four conditions while the fourth approached 2.0. Also, W3 was significantly affected by test length and IRT model at all four cut-off scores. W3 also produced significant interaction affects at all cut-off scores. It should be noted that with respect to the standard deviations of the four percentiles, W3 did seem to be more stable over replications, which is likely a consequence of the peaked distribution which characterized this index and the limited range of values for the index.

From the analysis of the non-aberrant response patterns, it would appear that ECIZ4 might be the most useful as an appropriateness index. That index produced a near normal distribution, was least affected by test length and IRT model, and was the most stable over replications. This would suggest that ECIZ4 may be used with more confidence than either Lz or W3 in practical test situations. Before making a final decision, the detection rate of aberrant patterns of the three indices must be examined.

Appropriateness Indices for Aberrant Response Patterns

In this section results of analysis of response patterns

which were modified to create various types and levels of aberrancy are presented and discussed. First, descriptive statistics are presented, then an analysis of detection rates is presented. Finally, the results are discussed and some possible explanations are proposed.

Descriptive Statistics for Aberrant Response Patterns

Tables 13, 14, and 15 present the mean and standard deviation for each index for the twenty-four combinations of test length, IRT model, type of aberrance, and level of aberrance. The mean values of the indices differ considerably from those produced in non-aberrant response patterns. The mean for Lz in non-aberrant response patterns was near 0 whereas the means for the aberrant response patterns ranged from -.67 to -3.07 over the twenty-four conditions. In the case of ECIZ4, which also produced a mean near 0 in non-aberrant response patterns, the range of values for aberrant response patterns was from .70 to 2.98. It is of interest that, in absolute value, the range of means was similar for both Lz and ECIZ4. For W3, which produced a mean near 1 for non-aberrant response patterns, the range of means was from 1.12 to 1.41. Interestingly, in aberrant response patterns, W3 appeared to produce the least deviation from the means of non-aberrant response patterns.

The standard deviation of the indices in the aberrant response pattern also produced higher values than those in non-aberrant response patterns. In the case of Lz and ECIZ4, the mean standard deviation over 50 replications was near 1.0 for the

Table 13

Mean and Standard Deviation of Lz for Aberrant Response Patterns

(N =4000)

		40 Item/2PLM		40 Item/3PLM		80 Item/2PLM		80 Item/3PLM	
Aberrance									
Type & Level		M	SD	M	SD	M	SD	M	SD
Spur. Low	10%	-1.38	1.55	-.84	1.47	-1.85	1.65	-1.23	1.61
	15%	-1.78	1.85	-.99	1.72	-2.45	2.04	-1.49	1.86
	30%	-2.19	2.41	-.67	1.81	-3.07	2.80	-1.26	2.14
Spur. High	10%	-1.21	1.51	-.72	1.11	-1.68	1.60	-1.03	1.12
	15%	-1.65	1.81	-1.02	1.25	-2.23	1.97	-1.40	1.38
	30%	-2.06	2.42	-1.41	1.82	-2.64	2.81	-1.82	2.18

non-aberrant response patterns. Over the 24 combinations of modified response patterns, the standard deviation of Lz ranged from 1.11 to 2.81. For ECIZ4 the standard deviation over the 24 combinations of modified response patterns ranged from 1.03 to 2.22. For W3 the range was .17 to .49, whereas the mean standard deviation in non-aberrant response patterns ranged from .14 to .23. The effects of the four independent variables are discussed below.

With respect to test length, the 80 item test produced more extreme values than the 40 item test for Lz and ECIZ4. Test

Table 14

Mean and Standard Deviation of ECIZ4 for Aberrant Response Patterns

(N = 4000)

		40 Item/2PLM		40 Item/3PLM		80 Item/2PLM		80 Item/3PLM	
Aberrance									
Type & Level		M	SD	M	SD	M	SD	M	SD
Spur. Low	10%	1.09	1.18	.70	1.13	1.58	1.23	1.10	1.28
	15%	1.46	1.42	.94	1.32	2.18	1.52	1.49	1.52
	30%	1.84	1.94	.83	1.38	2.98	2.22	1.60	1.71
Spur. High	10%	1.02	1.13	.74	1.02	1.45	1.19	1.15	1.03
	15%	1.43	1.33	1.04	1.08	1.99	1.43	1.59	1.17
	30%	1.89	1.86	1.43	1.39	2.59	2.13	2.02	1.60

length appeared to have little effect on W3. In addition, the standard deviation for Lz and ECIZ4 was consistently higher for the 80 item test as compared to the 40 item test. The opposite effect was observed for W3 although the differences in standard deviations were quite small.

The mean values of all three indices were more extreme for the 2PLM than for the 3PLM. The standard deviation of the 2PLM was also larger for each index in each condition than for the 3PLM. Of the three indices, W3 produced the least variability in standard deviation when the 2PLM was compared with the 3PLM.

Table 15

Mean and Standard Deviation of W3 for Aberrant Response Patterns

(N = 4000)

		40 Item/2PLM		40 Item/3PLM		80 Item/2PLM		80 Item/3PLM	
Aberrance									
Type & Level		M	SD	M	SD	M	SD	M	SD
Spur. Low	10%	1.24	.30	1.12	.24	1.24	.21	1.13	.19
	15%	1.33	.35	1.15	.30	1.33	.26	1.17	.23
	30%	1.41	.49	1.09	.37	1.41	.40	1.14	.29
Spur. High	10%	1.21	.30	1.13	.25	1.21	.22	1.13	.17
	15%	1.29	.36	1.19	.28	1.29	.26	1.17	.24
	30%	1.35	.49	1.23	.45	1.35	.41	1.22	.42

Results with respect to type of aberrance showed that, for Lz, spuriously low scores produced more extreme mean values than spuriously high scores in all conditions except for the higher levels of aberrance for the 3PLM. For ECIZ4 and W3, spuriously low scores tended to produce greater mean values for the 2PLM but spuriously high scores produced greater values for the 3PLM. These results further suggest interaction effects.

With respect to level of aberrance, increased aberrance typically produced more extreme mean scores for each index in each of the combinations of test length, IRT model, and type of

aberrance. For example, 15% aberrance tended to produce more extreme mean values than 10% aberrance, and 30% aberrance produced more extreme mean values than 15% aberrance. This was true for both spuriously low and spuriously high scores. The exception to this pattern of results was for the 3PLM spuriously low scores where 15% aberrance produced slightly more extreme mean values than 30% aberrance. For the 3PLM, mean values for both spuriously high and spuriously low were similar except for the 30% aberrance in the 40 item test, where spuriously high scores were somewhat greater than spuriously low scores.

In summary, with respect to the means and standard deviation of the indices in aberrant response patterns, Lz produced the greatest deviance from the mean for non-aberrant response patterns. W3 showed little variability within combinations of test length, IRT model, type of aberrance, and level of aberrance. Further, the 80 item test produced greater values than the 40 item test and the 2PLM greater values than the 3PLM. Spuriously low scores produced greater index values for the 2PLM and at 10% aberrance for the 3PLM. At higher levels of aberrance for the 3PLM, spuriously high scores tended to produce greater values. Higher levels of aberrance produced greater values in all conditions except for the 3PLM spuriously low conditions where 15% aberrance produced slightly higher index values than 30% aberrance for Lz and W3. For ECIZ4, higher levels of aberrance produced greater values in all cases but one.

Detection Rates

The detection rates for an appropriateness index are determined by examining the proportion of correct classification of aberrant response patterns at given false positive rates. In this study, four false positive rates were used; .01, .05, .10, and .25. Detection rates were computed for each of the twenty-four combinations of test length, IRT model, type of aberrance, and level of aberrance. The results are presented in Tables 16, 17, and 18 for each of Lz, ECIZ4, and W3 respectively. These tables present the marginal mean detection rates for each index for each test length, IRT model, and type of aberrance for each level of aberrance at each false positive rate. Detection rates for all of the 24 conditions for each index are included in Appendix A.

In Table 16 detection rates for Lz are shown. The 80 item test produced detection rates in a range from 21% to 71% across levels of aberrance and false positive rates. These were consistently higher than those of the 40 item test which ranged from 12% to 62%. Similarly, the 2PLM produced detection rates in a range of 22% to 72% which were consistently higher than the 3PLM which ranged from 11% to 60%. Spuriously low detection rates were higher than spuriously high detection at 10% and 15% aberrance at all false positive rates.

There seemed to be little difference between detection rates for the two types of aberrance. Contrary to expectations higher levels of aberrance did not systematically produce higher

Table 16

Percentage of Correct Classification in Aberrant Response Patterns for Lz

	False Positive Rates											
	.01			.05			.10			.25		
	Levels of Aberrance											
	10%	15%	30%	10%	15%	30%	10%	15%	30%	10%	15%	30%
40 item	12	20	26	27	35	37	37	44	48	62	61	57
80 item	21	32	37	38	45	47	48	53	52	66	71	65
2PLM	22	33	39	39	49	50	50	58	57	67	72	69
3PLM	11	19	24	26	34	34	36	43	41	55	60	54
Spur. high	14	23	31	30	40	42	40	49	49	60	66	62
Spur. low	19	28	31	36	43	42	45	51	48	63	66	60
2PLM/ spur .high	22	31	36	38	47	47	48	56	54	65	71	66
3PLM/ spur. high	8	16	27	22	22	37	33	42	45	54	62	59
2PLM/ spur. low	24	34	42	41	51	53	52	60	60	69	74	72
3PLM/ spur. low	15	22	21	30	36	31	39	43	37	57	59	48

detection rates. Further, at .01 false positive rate the combination of the 3PLM and spuriously low scores produced lower detection rates at higher levels of aberrance.

Detection rates for ECIZ4 are shown in Table 17. As in the

Table 17

Percentage of Correct Classification in Aberrant Response Patterns for ECIZ4

	False Positive Rates											
	.01			.05			.10			.25		
	Levels of Aberrance											
	10%	15%	30%	10%	15%	30%	10%	15%	30%	10%	15%	30%
40 item	9	17	24	22	33	36	32	43	45	55	64	65
80 item	17	31	37	36	49	52	48	60	64	57	78	80
2PLM	16	30	37	34	45	51	46	58	61	67	76	79
3PLM	9	18	24	24	34	37	35	45	48	57	66	67
Spur. high	11	22	30	27	40	45	38	52	58	63	73	77
Spur. low	14	26	30	30	41	43	41	50	52	61	68	68
2PLM/ spur.high	15	28	34	32	46	47	44	57	59	66	76	79
3PLM/ spur.high	8	17	28	23	35	42	34	48	57	60	71	76
2PLM/ spur.low	18	32	41	36	50	55	48	59	64	69	76	79
3PLM/ spur.low	11	20	20	25	33	31	35	42	60	54	61	57

case of Lz the 80 item test produced consistently higher detection rates than the 40 item test. For the 80 item test the range of values was from 17% to 78% while for the 40 item test the range was from 9% to 65%. The 2PLM produced detection rates

in a range from 16% to 76% which was consistently higher than the 3PLM where the range was from 9% to 67%. Detection rates for spuriously low scores ranged from 11% to 77% and tended to be higher than those for spuriously high scores at lower levels of aberrance and lower false positive rates. Detection rates for spuriously high scores were somewhat greater than for spuriously low scores at higher false positive rates and higher levels of aberrance. 15% aberrance produced higher detection rates than 10% aberrance in each condition, however, 30% did not produce higher detection rates than 15% aberrance in all cases. In general, detection rates were quite similar for both 15% and 30% aberrance although in some cases 15% aberrance produced higher detection rates than did 30% aberrance.

Results for detection rates for W3 are shown in Table 18. As with Lz and ECIZ4 the 80 item test with a range of from 12% to 72% produced higher detection rates than the 40 item test which ranged from 7% to 62%. The 2PLM with a range of from 13% to 73% produced higher detection rates than the 3PLM which ranged from 6% to 54%. Detection rates for spuriously low scores ranged from 11% to 66% and were higher than spuriously high scores except for the .05 false positive rate at 30% aberrance and the .25 false positive rate at 15% and 30% aberrance. As with Lz and ECIZ4, 15% aberrance produced higher detection rates than 10% aberrance in all conditions. However, 30% aberrance did not produce higher detection rates than 15% aberrance in each condition. In some conditions 15% aberrance produced slightly higher detection rates

Table 18

Percentage of Correct Classification in Aberrant Response Patterns for W3

	False Positive Rates											
	.01			.05			.10			.25		
	Levels of Aberrance											
	10%	15%	30%	10%	15%	30%	10%	15%	30%	10%	15%	30%
40 item	7	14	21	22	37	34	34	43	42	55	62	57
80 item	12	26	32	35	47	47	48	57	52	66	72	64
2PLM	13	27	33	35	47	46	47	58	54	66	73	67
3PLM	6	13	19	22	31	33	34	42	40	50	47	54
Spur. high	7	19	25	27	38	41	39	50	46	60	67	61
Spur. low	11	21	27	30	41	40	42	51	48	61	66	60
2PLM/ spur.high	10	26	29	33	45	42	45	56	50	64	70	63
3PLM/ spur.high	5	12	22	21	31	36	34	44	43	56	64	59
2PLM/ spur.low	15	28	37	37	50	50	50	60	59	68	75	70
3PLM/ spur.low	8	15	17	24	32	30	35	40	36	54	58	50

than 30% aberrance.

Comparisons of the effectiveness of the indices are presented in Tables 19, 20, and 21 for test length, IRT model, and type of aberrance respectively. As shown in Table 19 and as

Table 19

The Percentage of Correct Classifications of Three Indices for 40 Item and 80 Item Tests

		Levels of Aberrance					
		10%		15%		30%	
		40 Item	80 Item	40 Item	80 Item	40 Item	80 Item
False Positive	Index						
.01	Lz	12	21	20	32	26	37
	ECIZ4	9	17	17	31	24	37
	W3	7	12	14	26	21	32
.05	Lz	27	38	35	45	37	37
	ECIZ4	22	36	33	49	36	52
	W3	22	35	37	47	34	47
.10	Lz	37	48	44	53	48	52
	ECIZ4	32	48	43	60	45	64
	W3	34	48	43	57	42	52
.25	Lz	62	66	61	71	57	65
	ECIZ4	55	57	64	78	65	80
	W3	55	65	62	72	57	64

Table 20

The Percentage of Correct Classifications of Three Indices for the 2PLM and 3PLM

		Levels of Aberrance					
		10%		15%		30%	
		2PLM	3PLM	2PLM	3PLM	2PLM	3PLM
False Positive	Index						
.01	Lz	22	11	33	19	39	24
	ECIZ4	16	9	30	18	37	24
	W3	13	6	27	13	33	19
.05	Lz	39	26	49	34	50	34
	ECIZ4	34	24	48	34	51	37
	W3	35	22	47	31	46	33
.10	Lz	50	36	58	43	57	41
	ECIZ4	46	35	58	45	61	45
	W3	47	34	58	42	54	40
.25	Lz	67	55	72	60	69	54
	ECIZ4	67	57	76	66	79	67
	W3	66	55	73	47	67	54

Table 21

The Percentage of Correct Classifications of Three Indices for Spuriously High and Spuriously Low Scores

		Levels of Aberrance					
		10%		15%		30%	
		Sp.High	Sp.Low	Sp.High	Sp.Low	Sp.High	Sp.Low
False Positive	Index						
.01	Lz	14	19	23	28	31	31
	ECIZ4	11	14	22	26	30	30
	W3	7	11	19	21	25	27
.05	Lz	30	36	40	43	42	42
	ECIZ4	27	30	40	41	45	43
	W3	27	30	38	41	41	40
.10	Lz	40	45	49	51	49	48
	ECIZ4	38	41	52	50	58	52
	W3	39	42	50	51	46	48
.25	Lz	60	63	66	66	62	60
	ECIZ4	63	61	73	68	77	68
	W3	60	61	67	66	61	60

reported above, the 80 item test produced higher detection rates than the 40 item test for each index. Lz tended to produce the highest detection rates for both lengths of tests at 10% aberrance for all false positive rates. Lz also produced the highest detection rates at all levels of aberrance for .01 false positive rate. At higher false positive rates and higher levels of aberrance, ECIZ4 tended to produce the highest detection rates.

Tables 20 and 21 present the detection rates for IRT model and type of aberrance respectively. Generally, the results were similar for both independent variables insofar as Lz produced the highest detection rates at lower levels of aberrance and lower false positive rates. ECIZ4 produced higher detection rates for higher levels of aberrance and higher false positive rates. W3 did not produce the highest detection rate in any case although the detection rates were similar to Lz and ECIZ4 in some cases.

In summary, Lz produced the highest detection rates at .01 false positive rate at lower levels of aberrance and ECIZ4 generally produced the highest detection rates at 30% aberrance and at higher false positive rates. W3 produced the highest detection rate in only one condition although detection rates for W3 were similar to those of Lz and ECIZ4 in some conditions. The 80 item test produced higher detection rates than the 40 item test and the 2PLM higher rates than the 3PLM. There was little difference between detection rates for spuriously high scores or spuriously low scores. For type of aberrance there appeared to

be an interaction effect, such that at lower false positive rates, spuriously low scores produced higher rates and spuriously high scores produced higher rates at higher false positive. Generally, the combinations of the 3PLM with spuriously low scores produced lower detection rates than for other combinations of IRT model and type of aberrance. Also, increased aberrance tended to increase detection rate although in some conditions 30% aberrance was only equal to or less than that of 15% aberrance. Finally, the 80 item test tended to improve the detection rates for each index, especially in the case of ECIZ4.

A supplementary analysis carried out in this study was intended to examine the relationship between detection rates and examinee ability for each of the three indices. Six levels of ability (very low, low, low average, high average, high, very high) were established for the original generated abilities. Cut-off scores for the six levels of ability were -1.5, -.6, 0, .6, and 1.5. The detection rates in each group for each index at each false positive rate were then computed.

Six levels of estimated ability were also established and the percentages of correct classification for each rate were then computed. Similar proportions of examinees were classified in each of the six ability categories for both the original generated ability and the estimated ability. Results were produced for each of the 24 combinations of test length, IRT model, type of aberrance, and level of aberrance. As an example, the results for one of the analyses is shown in Appendix B.

Detection rates of the three indices appeared to be quite similar for both original ability and estimated ability. For example, at 30% spuriously high for Lz at .01 false positive rate, approximately 98% of the very low ability examinees were correctly detected for both of the original ability and the estimated ability. For ECIZ4, the detection rates were about 97% and 98% respectively and for W3 the detection rates were about 96% and 98% for the original and estimated ability categories. For 10% and 15% aberrance a similar pattern was found for each index over the two ability categories. Detection rates in middle range abilities were lower than those in either extremely high abilities under spuriously low conditions or extremely low abilities under spuriously high conditions. This suggests that the indices were not very effective in detecting unusual response patterns in average abilities. Since most tests have relatively many average scores, the efficacy of indices might be questioned for those examinees.

Discussion

In this section the effects of the four variables, test length, IRT model, type of aberrance and level of aberrance, on the detection rates of the three indices are discussed. With respect to the means and standard deviations of the indices in aberrant response patterns W3 produced the least amount of deviation from the means of the non-aberrant conditions. The lower standard deviations for W3 reflected the smaller range of values for W3. This is not surprising insofar as the standard

deviation of W3 showed the least variability in non-aberrant response patterns. The means and standard deviations of Lz showed most deviation from the mean in non-aberrant response patterns.

The effects of the independent variables were then examined. It was found that the 80 item test consistently produced higher detection rates than did the 40 item test. This might be explained as follows. Firstly, in examining cut-off scores based on non-aberrant patterns it was found that lower values were established for the 80 item test as compared to the 40 item test. This presumably happens because random errors can cause extreme index values to occur. In longer tests these random errors would tend to cancel out and result in lower cut-off scores; thus a higher percentage of aberrant response patterns would be detected. Secondly, for aberrant patterns the 80 item test would have twice as many aberrant responses as the 40 item test thus making them easier to detect.

With respect to IRT model, it was found that the 2PLM produced consistently higher detection rates than the 3PLM. For aberrant conditions, cut-offs for each false positive rate were less extreme for the 3PLM. Ordinarily it would seem that more aberrant patterns would be identified because of the lower cut-off score. Therefore explanations for higher detection rates under the 2PLM are speculative. It appears that using the 3PLM accounts for more of the randomness in item response generation procedures. This results in lower index values for the 3PLM

rather than the 2PLM. It may also be that the 2PLM less accurately estimates ability from modified response patterns; consequently, the appropriateness indices identified more modified response patterns as aberrant. Because ability was not re-estimated in the non-aberrant response patterns prior to computing the appropriateness indices, it may be that the response vector appears more aberrant for a particular ability. Consequently, the cut-off scores are higher than they may be expected.

Detection rates tended to be similar for both spuriously high and spuriously low scores. However, spuriously low scores produced slightly higher detection rate at lower false positive rates and lower levels of aberrance. This result is difficult to explain especially insofar as the differences are quite small (less than 5%) for most conditions. It was also found that for the 3PLM 40 item test condition, spuriously high scores tended to produce higher detection rates than spuriously low scores. This might be explained by the fact that with the 3PLM the "c" parameter has the effect of producing spuriously high scores. Thus more modified response patterns are identified.

The effect of levels of aberrance are as expected for most conditions insofar as 15% aberrance produced higher detection rates than 10% aberrance and 30% higher than 15%. However, under the 3PLM at higher false positive rates there was little difference between detection rates for 30% and 15% aberrance. This may be a result of the modification procedure where the

response patterns become saturated with predominantly 0's (for spuriously low) or 1's (for spuriously high). The re-estimated ability may therefore reflect the modified response pattern; thus the pattern does not appear aberrant with respect to the new ability estimate.

The results of the study also provided information on the relative effectiveness of the three indices under different test conditions. Lz appeared to be more effective than ECIZ4 or W3 at low levels of aberrance and low false positive rates. This may be explained by the fact that Lz had fairly extreme cut-offs at lower false positive rates; thus, the index may be somewhat more sensitive to aberrant scores than would ECIZ4. ECIZ4 may be moderated by the effect of the group response pattern. This would have the effect of producing slightly less extreme cut-off scores which in fact did occur. On the other hand Lz uses the probability of examinee's total response pattern and would be presumably more sensitive to extremes which may be unexpected for a particular ability. W3 is formed from the sum of residuals for each item which limits the effect of any one item on the size of the index. W3 was the most effective index in only one condition. One possible explanation is that W3 was found to produce a peaked distribution with a very restricted range. Further, Grosse and Wright (1988) observed that the weighted fit statistics may not be sensitive to outliers which would also affect detection rates.

Comparing the results of the detection rates found in this

study to those reported in other studies is somewhat problematic for two reasons. First, only Rudner (1983) and Drasgow, Levine, and McLaughlin (1987) have reported detection rates for standardized indices using simulation methods. Second, the specific conditions such as test length, IRT model, and level of aberrance differed from those used in this study. Nonetheless it is possible to provide some very general comparisons between the results of this study and those reported by Drasgow, Levine, and McLaughlin (1987). Rudner (1983) used only one false positive rate (.05) and only one of the indices (W3) used here and therefore the results of that study will not be examined further.

Results of this study showed that the range of detection rates from .05 to .10 false positive rates for 15% spuriously high was from 21% to 48% for Lz, from 23% to 57% for ECIZ4, and from 16% to 50% for W3. For examinees in mid-range abilities and 15% spuriously high, Drasgow, Levine, and McLaughlin (1987) report a range of from 18% to 48% for Lz, from 20% to 57% for ECIZ4, and from 15% to 53% for W3. For 15% spuriously low, results for this study showed detection rates for Lz to range from 27% to 49%, for ECIZ4 from 25% to 50%, and for W3 from 19% to 47%. Drasgow, Levine, and McLaughlin (1987) reported 16% to 49% for Lz, 11% to 42% for ECIZ4, and 11% to 44% for W3. It should be noted that those researchers used a slightly longer test, estimated item parameters, and homogeneous ability groups which presumably accounts for some of the differences.

For 30% spuriously low scores, Drasgow, Levine, and

McLaughlin (1987) reported results somewhat higher than those found here. At 30% spuriously high they found a range for Lz of 35% to 70% whereas for this study the results ranged from 26% to 42%. Similar differences were found for ECIZ4 and W3. The differences may be related to the fact that the detection rates were computed within a somewhat narrow range of high average abilities. Further, those researchers used estimated rather than generated item parameters. In terms of test length, Drasgow, Levine, and McLaughlin compared a 30 item arithmetic reasoning test with an 80 item combined test. The longer test produced systematically higher detection rates, results which correspond to those found here.

With respect to IRT model no comparative studies of the 2PLM and the 3PLM have been reported in the research, therefore it is not possible to offer comparisons with the results found here.

For type of aberrance some limited comparison may by made with the results reported by Drasgow, Levine, and McLaughlin (1987). Over false positive rates of from .01 to .10 they found spuriously high scores produced higher detection rates than did spuriously low scores in the average abilities. In the extreme ability ranges spuriously low scores produced higher detection rates. Results of this study found the detection rates overall quite similar between spuriously high and spuriously low scores. In some cases such as low false positive rates and low levels of aberrance spuriously low scores produced higher detection rates. In other cases spuriously high scores produced higher detection

rates. Because Drasgow, Levine, and McLaughlin (1987) used specific ranges of ability groups it is not possible to make meaningful comparisons for types of aberrance.

With respect to levels of aberrance Drasgow, Levine, and McLaughlin (1987) found 30% aberrance produced systematically higher detection rates than 15% aberrance. Results of this study indicate that, in general, 30% aberrance did not produce systematically higher detection rates than 15% aberrance. Again, because those researchers used high ability to produce spuriously low scores and low ability to produce spuriously high scores, the results may not be compared meaningfully.

Summary

This section describes briefly a summary of the results of the analysis of appropriateness indices first for non-aberrant response patterns and then for aberrant response patterns.

For non-aberrant response patterns ECIZ4 produced cut-off scores which were least affected by test length or IRT model. The distribution most closely approximated a normal distribution. Lz and W3 tended to produce more variation over test length and IRT model and were less stable over replications. Therefore it may be concluded that if the false positive rates for a test were to be based on a single sample, ECIZ4 may be the recommended appropriateness index.

For aberrant response patterns Lz produced higher detection rates than either ECIZ4 or W3 at lower levels of aberrance and

lower false positive rates. ECIZ4 produced somewhat higher detection rates for higher levels of aberrance and higher false positive rates. These findings suggest that, if appropriateness indices are being considered for test situations where it is important that one identify few false positives, Lz may be preferred. On the other hand, if false positive errors are more serious than failure to identify aberrant patterns, then one might wish to utilize ECIZ4. The distribution of W3 varied over replications and produced lower detection rates and therefore would not be recommended as an appropriateness index.

Finally, it is not clear whether the detection rates produced in this present study represent rates which may be of practical value in test interpretation, since such decisions would be related to the type and purpose of the test under consideration. Even Lz and ECIZ4, which produced some of the highest and most consistent results, did not provide detection rates as high as those reported in other research (Drasgow, Levine, & McLaughlin, 1987). This is undoubtedly due to the fact that cut-off scores for this study were determined by true ability whereas in the earlier study estimated ability was used. In the next chapter, the overall results of the study are reviewed and conclusions are presented with respect to those results.

CHAPTER 4

SUMMARY AND CONCLUSIONS

In this study two general problems in appropriateness measurement were identified. The first problem concerned the effect of tests length and IRT model on the characteristics of the distributions of three appropriateness indices in non-aberrant response patterns. The second general problem concerned the effectiveness of the three indices under different combinations of test length, IRT model, type of aberrance, and level of aberrance. In this chapter results of the study are summarized, some limitations are discussed, and suggestions for further research are presented.

Summary of Results

Following is a summary of the results of the study related to appropriateness indices in non-aberrant response patterns:

1. Lz and ECIZ4 produced an expected normal (0,1) distribution; W3 produced an expected mean of approximately one; the standard deviation was affected by IRT model and test length.

2. ECIZ4 produced the most nearly normal distribution which was also most stable over replications. ECIZ4 showed the smallest amount of skewness and little or no kurtosis; Lz showed a small amount of skewness and kurtosis; W3 showed some skewness and a very leptokurtic distribution.

3. Test length and IRT model did not affect the means of

the indices in non-aberrant response patterns.

4. Cut-off scores were established at four false positive rates. For each cut-off score, W3 was affected by test length, IRT model, and their interaction; Lz was affected by IRT model and test length; ECIZ4 was affected by IRT model at only one false positive rate.

In aberrant response patterns the following results were found with respect to the detection rates of the three indices:

1. There was not a large difference in detection rates among the three indices for most conditions. Overall detection rates ranged from 5% to 78% for Lz, from 5% to 88% for ECIZ4, and from 3% to 75% for W3.

2. Lz produced the highest detection rates for lower false positive rates and lower levels of aberrance; ECIZ4 produced the highest detection rates for higher false positive rates and higher levels of aberrance; W3 produced the highest detection rate in one of the 24 conditions. Lz may therefore be suggested as an appropriateness index if it is important to identify few false positive errors. If false positive errors are less serious than identifying aberrant response patterns then ECIZ4 may be preferred.

3. The 80 item test produced higher detection rates than the 40 item test.

4. The 2PLM produced higher detection rates than the 3PLM.

5. Spuriously low scores produced slightly higher detection rates than spuriously high scores although for the 3PLM at higher

false positive rates and higher levels of aberrance, spuriously high scores produced slightly higher detection rates.

6. 15% aberrance produced higher detection rates than 10% aberrance. 30% aberrance tended to produce higher detection rates than 15% aberrance; however in some cases, 15% aberrance produced detection rates equal to or higher than 30% aberrance.

Limitations of the Study

Although this study has provided a number of results with respect to the effect of test length and IRT model on appropriateness indices, interpretations of the results have a number of limitations. First, the data were generated by computer from specific ranges of item parameters and examinee ability. The item parameters were not drawn from a real test. Second, only two test lengths, 40 item and 80 item, were used and results may not be generalizable to other test lengths. It is not known whether any systematic relationship would exist over a range of test lengths such as 20, 30, 40, 60, 80, and 100 items. The study was also limited to the 3PLM and the 2PLM; another IRT model may have produced different results. Third, cut-off scores for the indices for non-aberrant response patterns were computed using the generated ability. If the ability had been estimated from the generated response vectors, there may have been a reduction in the values of the cut-off scores because the estimate of ability would more closely reflect the response pattern. A fourth limitation is that spuriously high and

spuriously low scores were analyzed separately. In real life a subject may produce both spuriously high and spuriously low scores which would presumably affect detection rates.

Suggestions for Further Research

Results of this study suggest that test length and IRT model affect the characteristics of the distributions of the indices. Further research is necessary to examine more precisely the effect of different test lengths on the indices. It would be of interest, for example, to know the effects on the indices of a very short test, say 15 or 20 items, or a very long test of more than 100 items. The effect of very short tests might be of particular interest to classroom teachers or where the cost of computer services is a concern.

A second area for further research concerns the use of other IRT models. It would be of interest to know the effect of the 1PLM because the 1PLM represents the simplest and least costly model to use and may be of benefit to test practitioners.

Third, it would be of interest to know the effect of simultaneously spuriously low and spuriously high scores of a subject. It is not really known if such a combination of types of aberrance may be detectable since they have not been studied when they appear in a single response pattern.

Fourth, this study used IRT based indices only. The distribution of group-dependent indices have been examined (Buxie, 19[illegible]); however, the effectiveness of those indices have

not been investigated extensively.

Fifth, examination of the effectiveness of appropriateness measurement in real data is not known. It would be of interest to know if the cut-off scores identified in this study would be of use in real life test conditions. The methodology for such an investigation may be problematic; however, the results may add information with respect to the validity of the indices.

Finally, there is a need to replicate this study to examine cut-off scores at other false positive rates.

REFERENCES

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

Blixt, S. L. and Dinero, T. E. (1985). An initial look at diagnoses based on Sato's caution index. Educational and Psychological Measurement, 45, 293 - 299.

Buxie, K. M. (1986). Effects of test difficulty, test length and simulated guessing on the distributional characteristics of aberrant response indices. Unpublished doctoral dissertation, University of Kansas, Kansas City.

Brennan, R. T. and Kane, M. T. (1977). An index of dependability for mastery tests. Journal of Educational Measurement, 14, 277 - 288.

Carlson, J. (1985). IBM Version of Datagen. [Computer program]. University of Ottawa.

Carroll, J. B., Meade, A., and Johnson, E. S. (1986). Test analysis with the person characteristic function. Manuscript submitted for publication.

Chatman, S. P. (1985). The relationship between response pattern aberrance and course performance in math placement. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Il.

Dixon, W. J. et al. (Eds.) (1985). BMDP statistical software manual. Berkeley: University of California.

Donlon, T. F. and Fischer, 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. (1987). Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72(1), 19 - 29.

Drasgow, F. and 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. and Levine, M. V. (1986). Optimal detection of certain forms of inappropriateness test scores. Applied Psychological Measurement, 10(1), 59 - 67.

Drasgow, F., Levine, M. V., and McLaughlin, M. E. (1987). Appropriateness measurement. (AFHRL - TP - 87 - 6). Manpower and Personnel Division, Brooks Air Force Base, Texas.

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

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

Frary, R. B. (1982). A comparison of person fit measures. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Garg, R. (1983). A comparison of the use of multiple matrix sampling and examinee sampling for test development. Unpublished doctoral dissertation, University of Ottawa, Ottawa.

Ghiselli, E. E. (1963). Moderating effects and differential reliability and validity. Journal of Applied Psychology, 47(2), 81 - 86.

Ghiselli, E. E. (1960). The prediction of predictability. Educational and Psychological Measurement, 20(1), 3 - 8.

Grosse, M. E. and Wright, B. D. (1988). Using latent trait and person fit statistics. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

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

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

Harnisch, D. L. and Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B. C.: Educational Research Institute of British Columbia.

Harnisch, D. L. and Torres, R. T. (1983). Techniques in detecting student errors: an investigation with a reading test. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Que.

Hulin, C. L., Drasgow, F., and Parsons, C. K. (1983). Item response theory: applications to psychological measurement. Homewood: Dow Jones - Irwin.

Jaeger, R. M. and Busch, J. C. (1986). The use and effect of caution indices in detecting aberrant patterns of standard setting recommendations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Ca.

Kane, M. T. and Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, 4(1), 105 - 126.

Koffler, S. L. (1983). A longitudinal analysis of curricular validity for a minimum competency testing program. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Que.

Levine, M. V. and Drasgow, F. (1982). Appropriateness measurement: review, critique, and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42 - 56.

Levine, M. V. and Drasgow, F. (1984). Performance envelopes and optimal appropriateness measurement. (Measurement Series 84 - 5). Model Based Laboratory. University of Illinois, Urbana.

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

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates.

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

Lumsden, J. (1978). Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 31, 19 - 26.

Linn, R. L. and Harnisch, D. L. (1981). Interactions between item content and group membership in achievement tests. Journal of Educational Measurement, 18, 109 - 118.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 3, 215 - 237.

Miller, D. (1986). Time allocation and patterns of response. Journal of Educational Measurement, 23, 147 - 156.

Nelson, R. B. and Chatman, S. P. (1986). The influence of guessing on measures of response aberrance when using the Rasch model. Paper presented at the annual meeting of the American Educational Research Association. San Francisco.

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

Parsons, C. K. (1983). The identification of people for whom job descriptive index scores are inappropriate. Organizational Behavior and Human Performance, 31, 365 - 393.

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

SAS Institute Inc. (1985). SAS user's guide: statistics, version 5 edition, (Computer program manual). Cary, NC: SAS Institute.

Schmitt, A. P. and Crocker, L. (1984). The relationship between test anxiety and person fit measures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Scheuneman, J. D. (1987). Item bias and individual differences. Manuscript submitted for publication.

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

Smith, R. M. and Wright, B. D. (1987). A new method of testing person fit in the Rasch model. Paper presented at the annual meeting of the American Educational Research

Association, Washington, D. C.

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

Tatsuoka, K. K. and Linn, R. L. (1983). Indices for detecting unusual response patterns: links between two general approaches and potential applications. Applied Psychological Measurement, 7(1), 81 - 96.

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

Tatsuoka, K. K. and Tatsuoka, M. M. (1982b). Standardized extended caution indices and comparison of their rule detection rates. (Research Report 82 - 4 - ONR). Urbana: University of Illinois, Computer Based Education Research Laboratory.

Tomsic, M. L. (1986). Stability of extended indices for standardized public school testing: a longitudinal study. Unpublished doctoral dissertation, University of Oregon, Eugene.

Trabin, T. E. and Weiss, D. J. (1979). The person response curve: fit of individuals to item characteristic curve model. (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 Zeitlinger, B. V.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97 - 115.

APPENDIX A

Detection Rates for Three Appropriateness Indices

Table A-1

Percentage of Correct Classification in Aberrant Response Patterns for Lz

			Levels of Aberrance											
			10%				15%				30%			
Test Length	IRT Model	Type of Ab	False Pos.Rate				False Pos.Rate				False Pos.Rate			
			.01	.05	.10	.25	.01	.05	.10	.25	.01	.05	.10	.25
40	2	High	15	31	41	60	24	40	50	66	31	43	50	64
40	3	High	5	17	27	48	10	26	36	57	22	33	41	56
80	2	High	27	44	54	70	38	54	62	75	41	51	57	68
80	3	High	10	26	38	60	21	38	48	66	31	41	48	61
40	2	Low	18	35	46	64	27	43	53	69	35	46	53	66
40	3	Low	11	25	33	52	16	29	37	53	15	25	31	43
80	2	Low	29	47	57	74	41	58	66	78	48	59	66	77
80	3	Low	19	35	44	61	27	42	49	65	26	36	42	54

Table A-2

Percentage of Correct Classification in Aberrant Response Patterns for ECIZ4

			Levels of Aberrance											
			10%				15%				30%			
Test Length	IRT Model	Type of Ab	False Pos.Rate				False Pos.Rate				False Pos.Rate			
			.99	.95	.90	.75	.99	.95	.90	.75	.99	.95	.90	.75
40	2	High	10	25	36	59	21	38	49	70	28	42	53	75
40	3	High	5	16	26	52	10	26	38	63	22	35	46	70
80	2	High	20	39	52	72	35	53	64	81	39	52	64	82
80	3	High	10	29	42	67	23	43	57	79	33	49	67	82
40	2	Low	12	28	39	61	23	40	49	68	31	44	52	69
40	3	Low	7	19	28	48	14	26	34	53	13	22	29	47
80	2	Low	23	43	56	76	41	59	68	83	50	66	76	80
80	3	Low	15	31	42	60	25	40	50	68	27	40	50	67

Table A-3

Percentage of Correct Classification in Aberrant Response Patterns for W3

Test	IRT Model	Type of Ab	Levels of Aberrance											
			10%				15%				30%			
			False Pos. Rate				False Pos.Rate				False Pos.Rate			
			.99	.95	.90	.75	.99	.95	.90	.75	.99	.95	.90	.75
40	2	High	9	25	37	58	18	37	48	65	24	38	47	62
40	3	High	3	16	27	49	7	24	37	60	16	31	40	58
80	2	High	11	41	53	69	33	53	63	75	33	45	52	64
80	3	High	6	26	40	62	16	38	50	68	27	40	46	60
40	2	Low	10	30	43	62	20	40	52	69	30	43	51	65
40	3	Low	4	18	28	50	10	26	35	52	12	24	30	44
80	2	Low	20	43	56	74	36	59	68	81	44	57	67	75
80	3	Low	11	29	41	58	19	37	47	63	22	35	42	55

APPENDIX B

Detection Rates in Original and Estimated Ability Groups

Table B-1

Percentage of Correct Classification for Original and Estimated Abilities for 2PLM, 40 Item, 30% Spuriously High

	N	Lz 01	ECI 99	W 99	Lz 05	ECI 95	W 95	Lz 10	ECI 90	W 90	Lz 25	ECI 75	W 75
Original Ability													
V.Low	286	98	97	97	99	99	99	100	100	100	100	100	100
Low	778	78	72	65	89	87	85	93	91	92	99	98	98
L.Ave	932	33	26	17	57	51	46	69	65	62	86	84	82
H.Ave	933	7	5	2	19	21	13	29	36	25	56	66	51
High	821	1	1	0	3	6	2	6	17	8	22	51	23
V.High	250	0	0	0	0	1	0	0	3	2	4	46	7
Estimated Ability													
V.Low	267	99	98	98	99	99	100	100	100	100	100	100	100
Low	761	80	78	69	92	90	89	95	93	93	99	98	98
L.Ave	1107	31	22	14	55	51	45	68	65	62	86	86	82
H.Ave	739	4	2	0	15	17	8	25	31	19	51	60	45
High	911	0	1	0	2	6	2	5	18	7	23	50	23
V.High	215	0	0	0	0	0	2	0	0	7	3	57	13

Cut-off Points for Ability Categories

Original Ability	Estimated Ability
-1.5	0
-0.6	0.9
0	1.6
0.6	2.0
1.5	3.0